Visual-Based Analysis of Classification Measures and their Properties for Class Imbalanced Problems

Dariusz Brzezinski*, Jerzy Stefanowski, Robert Susmaga, Izabela Szczęch

Institute of Computing Science, Poznan University of Technology,
ul. Piotrowo 2, 60–965 Poznan, Poland

Abstract

With a plethora of available classification performance measures, choosing the right metric for the right task requires careful thought. To make this decision in an informed manner, one should study and compare general properties of candidate measures. However, analysing measures with respect to complete ranges of their domain values is a difficult and challenging task. In this study, we attempt to support such analyses with a specialized visualization technique, which operates in a barycentric coordinate system using a 3D tetrahedron. Additionally, we adapt this technique to the context of imbalanced data and put forward a set of measure properties, which should be taken into account when examining a classification performance measure. As a result, we compare 22 popular measures and show important differences in their behaviour. Moreover, for parametric measures such as the F_β and $IBA_\alpha(G\text{-}mean)$, we analytically derive parameter thresholds that pinpoint the changes in measure properties. Finally, we provide an online visualization tool that can aid the analysis of measure variability throughout their entire domains.

Keywords: classification performance measures, visualization, barycentric system, class imbalance

1. Introduction

Classification is one of the most important machine learning tasks, commonly applied to many real-world problems. One of the crucial ingredients of this supervised learning task is the selection of a performance measure that allows the user to discern good classifiers from bad ones. An appropriate measure should support choosing the best classifier among several candidates and help tune its parameters. As a result, the selected performance measure is responsible for the optimization of the learning process [10].

Although researchers often focus on overall predictive accuracy, which promotes recognizing the highest number of instances of all target classes, the choice of the evaluation measure is not always a unique and simple decision. This is due to the fact that many practical classification problems require more sophisticated approaches to dealing with errors referring to particular subsets of instances. This realization has paved the way to proposals and analyses of many other performance measures.

Unfortunately, comparing several alternative classification measures and selecting the most appropriate one is not an easy task, even for experts constructing learning systems. In practice, crucial aspects of measures, such as atypical values they can take, their monotonicity, or their symmetry, are rarely taken into account during measure selection. As the understanding of measure properties is crucial for improving classification models and their learning process, we postulate the need for more research on analysing measures and developing methods that support researchers in this task.

*Tel.: +48 61 665 30 57

Email addresses: dariusz.brzezinski@cs.put.poznan.pl (Dariusz Brzezinski), jerzy.stefanowski@cs.put.poznan.pl (Jerzy Stefanowski), robert.susmaga@cs.put.poznan.pl (Robert Susmaga), izabela.szczech@cs.put.poznan.pl (Izabela Szczęch)

The analysis of measure properties may be done by examining a measure's definition. However, theoretical investigations are often very laborious and time consuming, especially when multi-dimensional aspects provided by the confusion matrices need to be taken into account. Owing to these difficulties, such an analysis could be alternatively carried out with visual techniques, in order to aid researches in finding and interpreting measure properties. Such visual-based insight is of utmost importance especially for classification tasks with complex example distributions, such as class imbalanced data.

In imbalanced data one of the target classes, called the minority class, contains much less examples than the remaining (majority) classes. Imbalanced data constitutes a great difficulty for standard learning algorithms, as classifiers tend to be biased towards the majority class and misclassify minority examples even though their correct recognition is usually more important [23, 30]. The prevalence of class imbalance in many practical tasks has led to the development of various methods for improving classifiers learning from skewed data [31, 44, 4]. In this context, much work has also been done in the field of classification measures. Since typical performance measures, such as classification accuracy, are not appropriate for imbalanced data [13, 36], several more relevant metrics have been considered. The most popular ones include *precision*, *recall* (*sensitivity*), *specificity*, and their aggregates, e.g. *G-mean* or F_1*-score*. These and other measures for imbalanced data are typically defined on the basis of confusion matrices summarizing the predictions of a binary classifier. Looking into related studies, one can notice that the number of such measures is relatively high and that each represents different aspects of classification performance, often leading to quite different interpretations [26]. This shows that there is no single measure that is the best choice in all situations. However, which measure is used in a given problem seems to be, to a large extent, dictated simply by the measure's popularity rather than a thorough discussion of its properties. That is why supporting visual analysis of performance measures is particularly important for class imbalanced problems and there is a need for new analysis methods.

In this paper, we put forward a new visualization technique for analysing entire domains of classification performance measures. The proposed visualization depicts all possible configurations of predictions in a confusion matrix, regardless of the used classifier. The method adapts an approach originally created for rule interestingness measures to the context of classification [39]. Contrary to existing performance measure visualizations, such as ROC space [14], the proposed method presents measures in a space which is defined directly on elements of the confusion matrix, is easily interpretable in 3D, and remains defined for all elements of the domain.

Acknowledging the need for systematic discussion on properties of performance measures, we also put forward ten properties which should be taken into account in the context of imbalanced data. The proposed properties characterize the behaviour of measures, reveal unexpected or atypical values, and can help researchers select measures suitable for a given learning problem.

Consequently, we consider 22 classification measures, chosen from the literature for their popularity and diversity, and analyse them with respect to the proposed properties. The analysis is performed using our visualization technique that allows to examine thoroughly the measures in complete ranges of their values. As a result, important differences in the measures' behaviour are highlighted, constituting a new theoretical contribution to research on class imbalance data and providing practical guidelines for selecting measures for particular classification problems. Finally, it is demonstrated that visualizations can also lead to analytical derivations of measure properties.

The main contributions of the paper are as follows:

- In Section 3, we present a technique for visualizing classification performance measures using the barycentric coordinate system and discuss its characteristics. Additionally, we present an online tool that implements the proposed technique and allows the analysis of both predefined as well as user-defined measures.
- In Section 4, we put forward ten properties, providing knowledge on the behaviour of classifier performance measures for imbalanced problems. The introduced properties involve maxima, minima, elements of symmetry, monotonicity, and undefined measure values.
- In Section 5, using the proposed visualization technique we analyse and compare 22 classification

measures with respect to the proposed properties. Moreover, we present a critical discussion on the applicability of measures with particular properties for imbalanced problems.

- In Section 6, we perform a set of case studies on how the proposed properties can be used to compare selected measures. More precisely, we study the differences between F_1-*score*, *G-mean*, *Mathews Correlation Coefficient*, and *Optimized Precision*, the effect of internal parametrization on the F_β measure, and external parametrization for $IBA_\alpha(G\text{-}mean)$. Apart from visual inspection, we analytically derive threshold parameter values for selected measures.
- In Section 7, we discuss the most important issues in analysing classification performance measures and draw lines of potential further investigations.

2. Related Works

2.1. Classifier performance measures

Classifiers can be assessed in many aspects, such as their predictive ability, training time, memory usage, model complexity, interpretability, or other criteria [27]. In this paper, we consider predictive performance only and focus on measures that evaluate crisp binary classifier predictions; measures specific to only rankers or probabilistic classifiers are out of the scope of this study. Furthermore, we concentrate mainly on measures which take into account the binary class imbalance problem.

As discussed in [23], when dealing with imbalanced data measures should focus on the minority class. Such measures are defined as functions of the confusion matrix for two-class problems, with the minority class typically referred to as *positive* (P), while the remaining majority class as *negative* (N) [27, 22] (multiple non-positive classes, if present, are usually aggregated into one).

Table 1: Confusion matrix for two-class classification

Predicted / Actual	Positive	Negative	total
Positive	TP	FN	P
Negative	FP	TN	N
total	$\widehat{P}$	$\widehat{N}$	n

Table 1 illustrates a two-class confusion matrix, which may be regarded as a special case of a contingency table that can be multi-dimensional in general. The TP (*True Positive*) and TN (*True Negative*) entries denote the number of examples classified correctly by the classifier as positive and negative, while the FN (*False Negative*) and FP (*False Positive*) indicate the number of misclassified positive and negative examples, respectively. Based on these values, the most common performance measures are defined as:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$precision = \frac{TP}{TP + FP} \quad (2)$$

$$specificity = \frac{TN}{FP + TN} \quad (3)$$

$$sensitivity\ (recall) = \frac{TP}{TP + FN} \quad (4)$$

Many other classification performance measures were proposed based on values from the confusion matrix; for their reviews see [23, 27, 25, 21, 2]. In this study, we analyse the properties of 22 measures, listed and defined in the supplementary material.[1] Below, we highlight four measures, chosen for diversity of their characteristics, which we will analyse and compare in more detail in Sections 5 and 6:

[1] https://dabrze.shinyapps.io/Tetrahedron/

$$F_\beta = \frac{(1+\beta)\cdot \mathit{precision} \cdot \mathit{recall}}{\beta \cdot \mathit{precision} + \mathit{recall}}, \text{ where } \beta \geq 0, \tag{5}$$

$$G\text{-}mean = \sqrt{\mathit{sensitivity} \cdot \mathit{specificity}}, \tag{6}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{\widehat{P} \cdot P \cdot N \cdot \widehat{N}}}, \tag{7}$$

$$OP = \mathit{accuracy} - \frac{|\mathit{specificity} - \mathit{sensitivity}|}{\mathit{specificity} + \mathit{sensitivity}}. \tag{8}$$

F_β combines *precision* and *recall* as a weighted harmonic mean, with the β parameter as their relative weight. Commonly $\beta = 1$ and then the measure is referred to as F_1*-score*. *G-mean* [28] is the geometric mean of *sensitivity* and *specificity*, which takes into account the relative balance of recognition of both positive and negative classes. The *Matthews Correlation Coefficient* (*MCC*) expresses a correlation between the actual and predicted classification and returns a value between −1 (total disagreement) and +1 (perfect agreement). We highlight *MCC* in our study since it was considered by some authors as one of the recommended measures for imbalanced data [2, 3]. *Optimized precision* (*OP*) combines *sensitivity* and *specificity* in a more complex way, also producing values in the $[-1, +1]$ range [37].

Apart from these "closed-formula" measures, we shall also analyse in more detail a representative of what may be thought of as "open-formula" measures, in this case $IBA_\alpha(M)$. This particular measure-wrapper is aimed at applying more weight to minority class predictions in a given measure M, according to a user-defined parameter α [17, 18].

These and other measures were compared in such surveys as e.g. [23, 21, 22, 2], however usually with the aim of discussing the main differences in their definitions. Additionally, the F_1*-score* was thoroughly analysed by Powers [35] who claimed that some of its properties, such as focusing only on the minority class and assuming that actual and predicted distributions are identical, may be critical flaws. Another theoretical study showed that aggregating *sensitivity* and *specificity* presented more suitable behaviour than aggregating *precision* and *recall* [25]. Nevertheless, theoretical analyses of measures with respect to complete ranges of domain values are very laborious and have been done only for a few classifier performance measures.

2.2. Visualization of measures

In this paper, we focus on visualizing measures defined on a binary confusion matrix. We note that this should not be confused with visualizations of classifier performance, e.g. using ROC graphs [12], precision-recall curves [9], lift charts [34], or other attempts to graphically present experimental comparisons of classifiers [41, 1, 7, 11, 24]. Our intention is to study general properties of measures rather than visualize the predictive performance of a classifier on a given dataset.

The 3D visualizations of 2 × 2 sum-constrained matrices, applicable in particular to confusion matrices, have already been considered in different papers. Below, we recapitulate shortly three approaches, which bear some relation to the (regular) tetrahedron visualization used throughout this paper [29, 8, 14].

Le Bras et al. [29] introduce a system of 3D spaces (referred to as the *Formal Framework*), in which the contents of sum-constrained 2 × 2 matrices can be represented. Because of the three actual degrees of freedom of a sum-constrained 2 × 2 matrix, domains consisting of three variables are required and sufficient to express the matrix entries. However, the choice of a particular domain, with three particular variables, may vary depending on the application at hand.

While the representations with three variables might be used to produce 3D visualizations of measures, the paper of Le Bras et al. [29] does not exploit this fact in too much a detail, as its focus lies elsewhere. The authors introduce three very particular, application-driven, 3D domains referred to as: *confidence*, *examples* and *counterexamples*. In its central part, the paper recalls 38 measures related to association rules and defines them consistently in terms of the matrix entries, as well as in terms of the three proposed domains. This allows for conducting dedicated analyses of the measures (e.g. expressing the Piatetsky-Shapiro recommendations [33] in the *examples* domain), with the main objective of identifying measures most relevant to association rule pruning. The introduced and in detail examined properties include: all-monotonicity, generalized universal existential upward closure, and opti-monotonicity [29].

As far as the tetrahedron-based visualization is concerned, the *examples* and *counterexamples* 3D spaces introduced in [29] assume the shapes of tetrahedra. However, contrary to the approach presented in our paper, the domains are designed for analysing rule interestingness measures. Moreover, the tetrahedra of the *examples* and *counterexamples* domains are irregular, since these domains are assumed to have two orthogonal variables each, implying shapes with two orthogonal edges incident with one vertex, a feature unattainable in the regular tetrahedron.

Celotto [8] has introduced 2D visualization spaces that are very natural to the considered measures, i.e., Bayesian confirmation measures. The primary space, suitably referred to as the *confirmation space*, consists of: $P(H|E)$ (x-axis) and $P(H)$ (y-axis). As noted within the paper, the 2D representation of 2×2 sum-constrained matrices is incomplete, and thus aptly called a *fingerprint* of the measure. The incompleteness results from the fact that the fingerprint changes as some third parameter which defines the third dimension, in this case chosen to be $P(E)$, is varied.

The confirmation space is initially set side by side with its analogue, denoted as *dual confirmation space*, and another 2D space, i.e. the ROC space, which consists of false positive rate $fpr = FP/N$ (x-axis) and true positive rate $tpr = TP/P$ (y-axis). However, because confirmation measures remain the main focus of the study of Celotto [8], presented analyses are basically confined to the confirmation space and its dual, which are used to analyse 19 measures. The measure analyses, principally concerned with identifying measures most relevant to classification rule pruning, include visualizations of some ordinal equivalence aspects and a multitude of symmetry aspects. The latter also include visually-assisted design and synthesis of measures possessing desired symmetries.

The 2D confirmation spaces introduced by Celotto correspond to rectangular cross-sections of the 3D tetrahedron presented in this paper. However, contrary to the presented approach, in [8] these originally non-independent variables are presented as orthogonal and with unified ranges, which thus requires some amount of orthonormalization.

Flach [14] mentions several possible definitions of variables suitable for 3D visualizations of 2×2 confusion matrices, but focuses primarily on *3D ROC space*, a generalization of traditional 2D ROC space [27]. The 3D ROC space consists of the false positive rate $fpr = FP/N$ (x-axis) and true positive rate $tpr = TP/P$ (y-axis), which basically constitute traditional ROC space, together with the frequency of positives $pos = P/n$ (z-axis). This choice had been dictated by the general topic of the paper, which was the analysis of classifier performance measures and their behaviour in ROC spaces. Notice that the three variables are selected so that the resulting XY-plane hosts the ROC space, while the third co-ordinate varies with the actual class distribution. In result, the 3D ROC space is thus a collection of stacked-up ROC spaces, with the z-coordinate corresponding to the proportion of the positive class. Owing to the variable mutual orthogonality and similar ranges ($[0, 1]$ for x and y and $(0, 1)$ for z) the total domain shape is thus a $[0, 1] \times [0, 1] \times (0, 1)$ pseudo-cube, i.e. a cube with both the lowermost layer, corresponding to $FN + TP = 0$, and the uppermost layer, corresponding to $FP + TN = 0$, removed. Similar techniques have been used to analyse rule quality measures. The most well known are coverage spaces [16], which plot the number of positive training examples and negative ones covered by the rule in the given data.

In the cited study [14], Flach combines the proportion of classes with misclassification costs, generally referred to as skew, and focuses on analysing 8 selected measures in terms of sensitivity to skew. The considered key notions involve: skew-equivalence and weak/strong skew-insensitivity of the measures. The stacked 2D spaces considered by Flach [14] basically correspond to the rectangle-shaped cross-sections of the tetrahedron presented in this paper. However, ROC spaces are presented in square form, which requires some amount of orthogonal rescaling compared to the approach presented in this paper. Furthermore, contrary to the visualization technique introduced in this paper, 3D ROC space remains undefined for confusion matrices with $FN + TP = 0$ or $FP + TN = 0$.

3. The barycentric visualization technique

As presented in Table 1, a confusion matrix for binary classification consists of four entries: TP, FP, FN, TN. However, for a dataset of n examples these four entries are constrained, as $n = TP + FP + FN + TN$. Therefore, for a given constant n, any three values in the confusion matrix uniquely define the fourth one. This property allows to visualize any classification performance measure based on the two-class confusion matrix using a 4D barycentric coordinate system.

In *barycentric coordinate systems* [43] point locations are generally expressed with relation to pre-defined points rather than as combinations of fixed unit vectors (as is the case with the Cartesian coordinate system), which lends them more generality. In its originally mechanical conception [32], the system states that if each of the pre-defined points is assigned (positive) mass, a given point is located at the mass centre (or barycentre). On the grounds of the resulting generality, barycentric coordinate systems find extensive applications in computational geometry in general and geometric modelling in particular [15].

In the considered barycentric coordinate system point locations are specified relatively to vertices of a tetrahedron, where each dimension is represented as one of the four vertices. Choosing vectors that represent TP, FP, FN, TN as vertices of a regular tetrahedron in a 3D space, one arrives at a barycentric coordinate system as in Fig. 1.

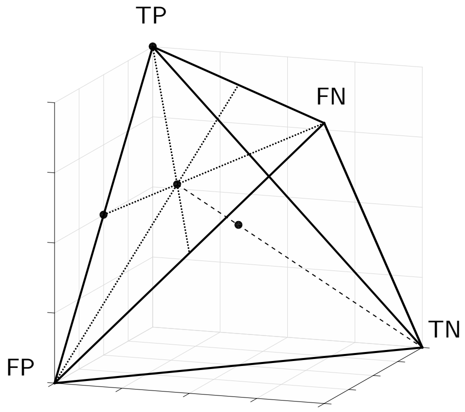


Fig. 1. A skeleton visualization of the tetrahedron with four exemplary points

In this system, every confusion matrix $\left[\begin{smallmatrix} TP & FN \\ FP & TN \end{smallmatrix}\right]$ is represented as a point of the tetrahedron. Let us illustrate this fact with a few examples. Fig. 1 shows a skeleton of a tetrahedron with 4 exemplary points:

- one located in vertex TP, which represents $\left[\begin{smallmatrix} n & 0 \\ 0 & 0 \end{smallmatrix}\right]$,
- one located in the middle of edge TP–FP, which represents $\left[\begin{smallmatrix} n/2 & 0 \\ n/2 & 0 \end{smallmatrix}\right]$,
- one located in the middle of face △TP–FP–FN, which represents $\left[\begin{smallmatrix} n/3 & n/3 \\ n/3 & 0 \end{smallmatrix}\right]$,
- one located in the middle of the tetrahedron, which represents $\left[\begin{smallmatrix} n/4 & n/4 \\ n/4 & n/4 \end{smallmatrix}\right]$.

One way of understanding this representation is to imagine a point in the tetrahedron as the centre of mass of the examples in a confusion matrix. If all n examples are true positives, then the entire mass of the predictions is at TP and the point coincides with vertex TP. If all examples are false negatives, the point lies on vertex FN, etc. Generally, whenever $a > b$ ($a, b \in \{TP, FN, FP, TN\}$) then the point is closer to the vertex corresponding to a rather than b.

The barycentric coordinate system makes it possible to depict the originally 4D data (two-class confusion matrices) as points in 3D. Moreover, as in [39, 40], an additional variable based on the depicted four values may be rendered as colour. Although any colour map can be used, in the following paragraphs we utilize the map, often used in cartography, shown in Fig. 2: dark blue — minimum values, dark brown — maximum values. Areas of the same colour signify then the same values of the variable. The shape of such areas is determined by the nature of the visualized variable and usually occurs as lines in 2D (isolines) and surfaces in 3D (isosurfaces). Undefined values of the measures will be rendered in magenta, i.e., a colour not occurring in the map.

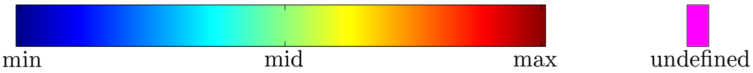


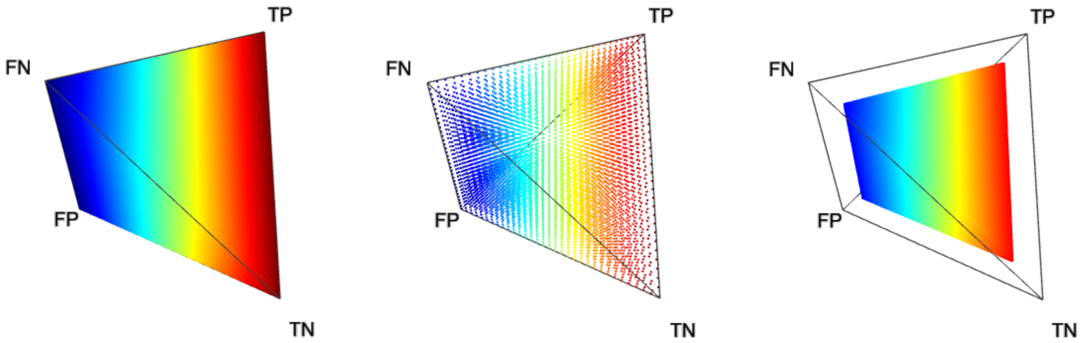
Fig. 2. Colour map used to depict measure values

Here, we adapt this procedure to colour-code the values of classification performance measures, which remain the principal focus of this paper. In this respect, the presented approach is different from [39] and [40], in which Bayesian confirmation measures were mainly addressed. In particular, this paper introduces and discusses those aspects of the tetrahedron-based visualization that are especially useful for the analysis of classification performance measures.

The described visualization technique has been implemented as an interactive web application, available at: https://dabrze.shinyapps.io/Tetrahedron/. The application can visualize 86 predefined 4D measures, including the 22 classification performance measures described further. The user can also visualize custom measures by providing their formulae. For the remainder of the paper, the reader is encouraged to use this tool to interactively analyse the described properties of various classification measures.

Note that in order to reveal all possible measure behaviours, the visualization technique needs to depict all possible combinations of confusion matrix entries. Thus, the visualizations are rendered using predefined data consisting of all such confusion matrices for a given n, and not using data provided by a user for a particular problem. It is due to the fact that data from a particular task might not cover the whole measure domain, leaving thus some of its parts not inspected, and ripping the knowledge about the analysed measure off the necessary generality.

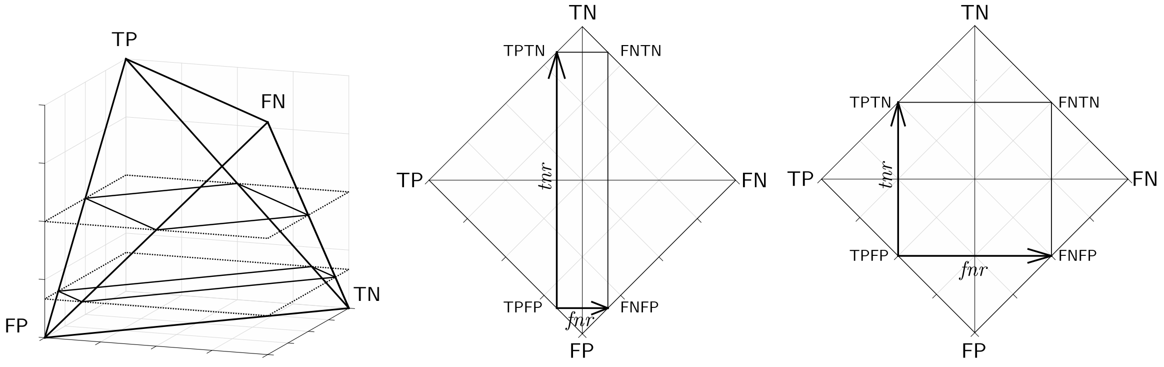
Since classification *accuracy* is one of the simplest and most often used performance measures, let us use it for an exemplary visualization in Fig. 3. Its values range from 0 to 1, and there are no undefined ones. One can notice that confusion matrices with a high number of *FP* and *FN* result in low *accuracy*, whereas high *TP* and *TN* yield high *accuracy*. The visualization in Fig. 3a is only partially comprehensive, as it merely shows the externals of the tetrahedron which correspond to very specific confusion matrices. However, both external as well as internal areas can be shown, e.g. by padding tetrahedron points (Fig. 3b), using "under the skin" views (Fig. 3c) or performing cross-sections (Fig. 4).



(a) External view **(b)** Point padding **(c)** Internal view

Fig. 3. Visualizations of classification *accuracy*

The indicated cross-sections are of particular interest in the context of analysing measures for class imbalance problems. Notice that traversing the tetrahedron alongside the vertical axis (up-down in Fig. 4a)



(a) External view **(b)** $P/n = 1/6$ **(c)** $P/n = 1/2$

Fig. 4. Skeleton visualizations of the tetrahedron and top-down view depictions of rectangular cross-sections for two selected values of positive class rate (P/n)

corresponds to changing the proportions between sums $TP + FN = P$ and $FP + TN = N$, which specify the cardinalities of the actual classes. If $P = N$, then a situation of balanced classes is reproduced; otherwise the classes are imbalanced.

How a measure behaves for a particular class proportion may be visualized by producing a cross-section of the tetrahedron with a horizontal plane that cuts its vertical height. Fig. 4b and 4c show the two cross-sections visible in Fig. 4a, one at $P/n = 1/6$ (positive class as the minority class) and one at $P/n = 1/2$ (class balance), as seen from above the tetrahedron. Cutting the shape with a horizontal plane at $P/n = 1/6$ produces the lower, rectangular cross-section (Fig. 4b), while at $P/n = 1/2$ — the upper, square one (Fig. 4c). In their corresponding figures, the cross-sections are oriented so that their sides incident with face △TP–FP–FN of the tetrahedron are positioned at the bottom, while those incident with face △TP–FN–TN — at the top. It is additionally worth noting that the proportion of the rectangle's side lengths follows that of the class cardinalities, P (the horizontal side) and N (the vertical side).

Accordingly to the notation of the vertices of the tetrahedron, the sides and vertices of a cross-section rectangle are labelled as follows:

- sides: $\overline{\mathsf{TP}}$ (left), $\overline{\mathsf{TN}}$ (upper), $\overline{\mathsf{FN}}$ (right), $\overline{\mathsf{FP}}$ (lower),
- vertices: TPTN (upper-left), FNTN (upper-right), FNFP (lower-right), TPFP (lower-left).

The two axes, *fnr* and *tnr*, of the 2D space in which all cross-sections are represented (including those for $P/n = 1/6$ and $P/n = 1/2$), correspond to the false negative rate, $fnr = FN/(FN + TP) = 1-$ *recall*, and the true negative rate, $tnr = TN/(TN + FP) =$ *specificity*. The orientation of the axes results from the fact that traversing the rectangle left-to-right corresponds to increasing *fnr* from 0 to 1, whereas traversing the rectangle down-up corresponds to increasing *tnr* from 0 to 1. The resulting 2D space of the presented cross-section is thus an analogue of 2D ROC space, where, somewhat reversely, the false positive rate, $fpr = FP/(FP + TN) = 1-$ *specificity*, and the true positive rate, $tpr = TP/(FP + TP) =$ *recall*, are used as x and y axes, respectively.

The presented rectangular cross-sections and 2D ROC space constitute the same, though seen from different angles, cross-sections of the tetrahedron. However, contrary to 3D ROC space [14], the presented technique does not involve any non-linear transformations of the elements of the confusion matrix and remains defined for all elements of the domain. Furthermore, because the proposed barycentric coordinates directly correspond to elements of the confusion matrix, the visualization is easily interpretable also in 3D, which helps analysing the whole range of possible domain values.

In the following sections, we demonstrate the usage of the visualization technique in analyses of classifier performance measures for imbalanced data. The technique, including the cross-sections, was particularly used to visualize several properties of the measures.

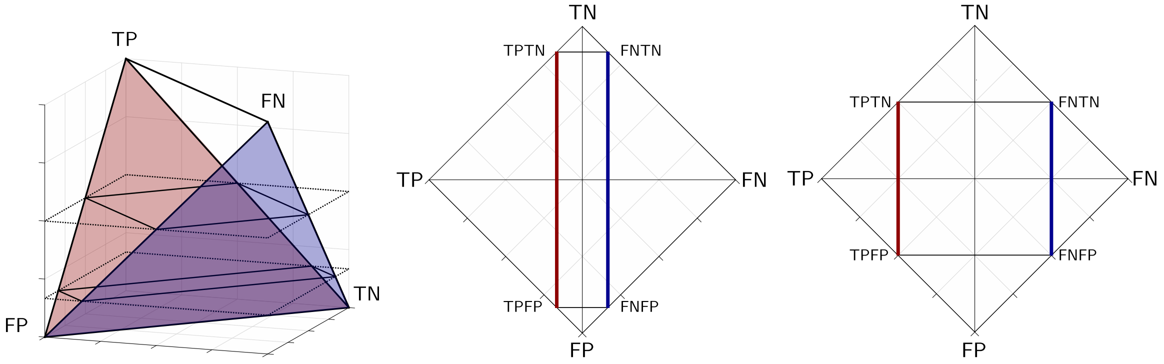
4. Properties of Measures for Imbalanced Data

With a visualization technique at hand, it is much easier to define and interpret potentially desirable measure properties. The properties should aid researchers in comparing various classification measures, examining differences in changes of their values with respect to possible predictions, noticing unusual or unexpected values, and selecting measures suitable for a given task. The analysis of such properties could raise a more systematic discussion on the applicability of existing measures in certain domains and guide the proposal of new measures.

All these issues are particularly important for class imbalance problems, where the number of available measures is relatively high. Each of them represents different aspects of classifier performance and expert preferences, making their interpretation and analysis a non-trivial task. Therefore, in this section we put forward ten properties, which capture desired characteristics of classifier performance measures in the context of imbalanced data. We explain the details of each property using the introduced visualization technique and discuss their usefulness in practice.

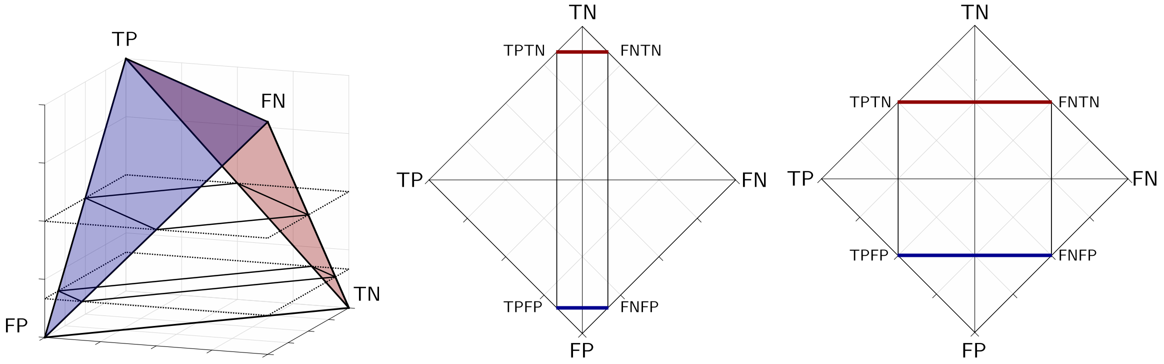
To better explain the properties, we recall that the interpretation of the rectangular cross-section discussed in Section 3 is as follows.

- side $\overline{TP}$ / $\overline{FN}$: full/null recognition of the positive (minority) class (Fig. 5),
- side $\overline{TN}$ / $\overline{FP}$: full/null recognition of the negative (majority) class (Fig. 6).



(a) External view **(b)** $P/n = 1/6$ **(c)** $P/n = 1/2$

Fig. 5. Illustration of full/null recognition of the positive class



(a) External view **(b)** $P/n = 1/6$ **(c)** $P/n = 1/2$

Fig. 6. Illustration of full/null recognition of the negative class

In this context, we postulate to analyse classifier performance measures with respect to ten properties:

TPTN_{max}: vertex TPTN maximal value,

$\overline{\mathsf{FN}}_{min}$: side $\overline{\mathsf{FN}}$ minimal value,

$\overline{\mathsf{FP}}_{min}$: side $\overline{\mathsf{FP}}$ minimal value,

$TP_{\nearrow}$: horizontal lines weakly monotonic value growth (from $\overline{\mathsf{FN}}$ to $\overline{\mathsf{TP}}$),

$TN_{\nearrow}$: vertical lines weakly monotonic value growth (from $\overline{\mathsf{FP}}$ to $\overline{\mathsf{TN}}$),

$\overline{\mathsf{TN}}_{\neq max}$: side $\overline{\mathsf{TN}}$ less than maximal value except for vertex TPTN,

$\overline{\mathsf{TP}}_{\neq max}$: side $\overline{\mathsf{TP}}$ less than maximal value except for vertex TPTN,

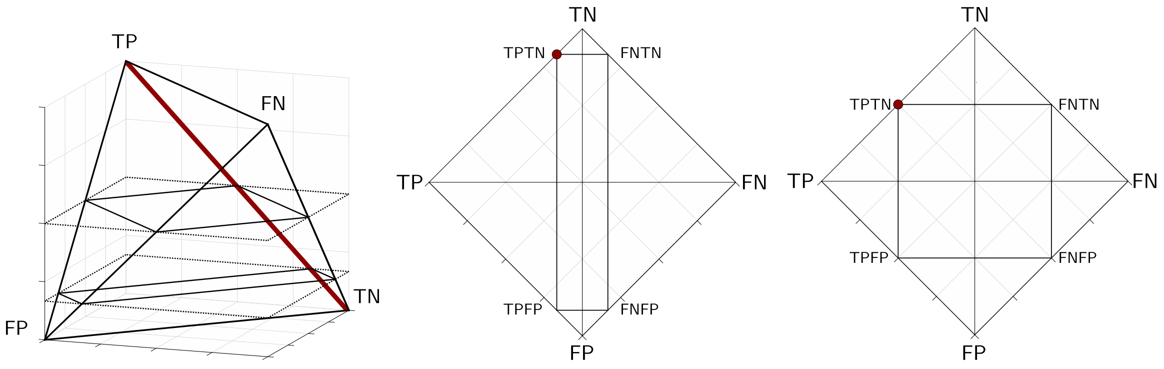
ACE: for any two corresponding points on sides $\overline{\mathsf{TP}}$ and $\overline{\mathsf{TN}}$ (e.g. middle points) the value on side $\overline{\mathsf{TP}}$ is greater or equal to that on $\overline{\mathsf{TN}}$,

ACH: values invariant under exchange of TP with TN and FN with FP,

UnDefs: the existence (and the location) of undefined values.

If present, undefined measure values are excluded from the above considerations, except for the last property, which is directly concerned with those values. Similarly, all but the last two properties are analysed only for 'non-degenerated' rectangular cross-sections, i.e. cross-sections corresponding to $P > 0$ and $N > 0$. On the other hand, the 'degenerated' cross-section, i.e. cross-sections that result in rectangles of either zero breadth or zero width, are taken into account in the ACH and *UnDefs* properties.

Notice that although the properties can be analysed using rectangular cross-sections, the considerations naturally extend from 2D in the rectangles to 3D in the tetrahedron when all feasible cross-sections of the considered type are taken into account. For example, points TPTN of all rectangles collectively form edge TP–TN of the tetrahedron (Fig. 7), whereas sides $\overline{\mathsf{FN}}$ of all rectangles collectively form face $\triangle\mathsf{FP}$–FN–TN of the tetrahedron (Fig. 8). In result, these properties might be defined and examined in 3D, thus capturing the generally multi-dimensional nature of the measures with which they are concerned. This aspect further emphasizes the usefulness of the introduced visualization.



(a) External view **(b)** $P/n = 1/6$ **(c)** $P/n = 1/2$

Fig. 7. Illustration of property TPTN_{max}

The presented properties may be regarded as a basic 'check-list', providing knowledge on the behaviour of classifier performance measures for imbalanced data. Below, let us discuss each of them in detail.

Recall that the analysed measures are functions of $TP \geq 0$, $FN \geq 0$, $FP \geq 0$ and $TN \geq 0$, $TP + FN + FP + TN = n$, which constitute the elements of the confusion matrix $\left[\begin{smallmatrix} TP & FN \\ FP & TN \end{smallmatrix}\right]$ (see Table 1). In this context, $f(\left[\begin{smallmatrix} TP & FN \\ FP & TN \end{smallmatrix}\right])$ denotes the value of any of the considered classification performance measures.

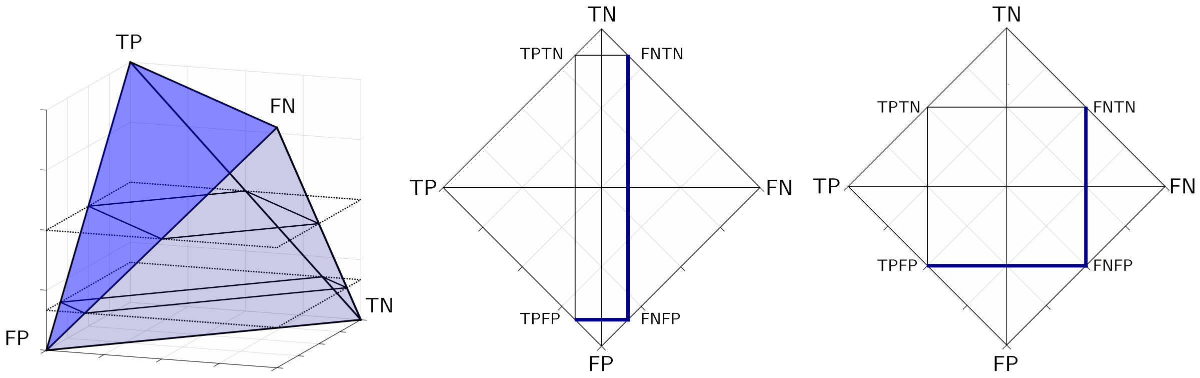
Property TPTN_{max} ensures that perfect predictions of both classes always render the best measure value (Fig. 7). Notice that vertex TPTN, being the common part of both side $\overline{\mathsf{TP}}$ and side $\overline{\mathsf{TN}}$, is actually the only point of full recognition of both the positive and the negative class. Because TPTN corresponds to $\left[\begin{smallmatrix} P & 0 \\ 0 & N \end{smallmatrix}\right]$, this implies:

$$\mathsf{TPTN}_{max} : f(\left[\begin{smallmatrix} P & 0 \\ 0 & N \end{smallmatrix}\right]) = max.$$

Properties $\overline{\mathsf{FN}}_{min}$ and $\overline{\mathsf{FP}}_{min}$ state that not recognizing one of the classes should correspond to the worst possible measure value (Fig. 8). They formalize basic expectations towards good classifier performance, which are often expressed in the literature on class imbalances. In many situations, the standard classifiers may completely fail to correctly classify examples from the minority class while being sufficiently good at recognizing the majority class [5]. On the other hand, attempts to improve classification of the minority class examples should not deteriorate the recognition of the majority class [20]. In binary classification, a null recognition of any of the two classes is certainly insufficient, thus, it is naturally required that measures should obtain minimal values on sides $\overline{\mathsf{FN}}$ and $\overline{\mathsf{FP}}$. This boils down to:

$$\overline{\mathsf{FN}}_{min} : f(\left[\begin{smallmatrix} 0 & P \\ FP & TN \end{smallmatrix}\right]) = min,$$

$$\overline{\mathsf{FP}}_{min} : f(\left[\begin{smallmatrix} TP & FN \\ N & 0 \end{smallmatrix}\right]) = min.$$



(a) External view **(b)** $P/n = 1/6$ **(c)** $P/n = 1/2$

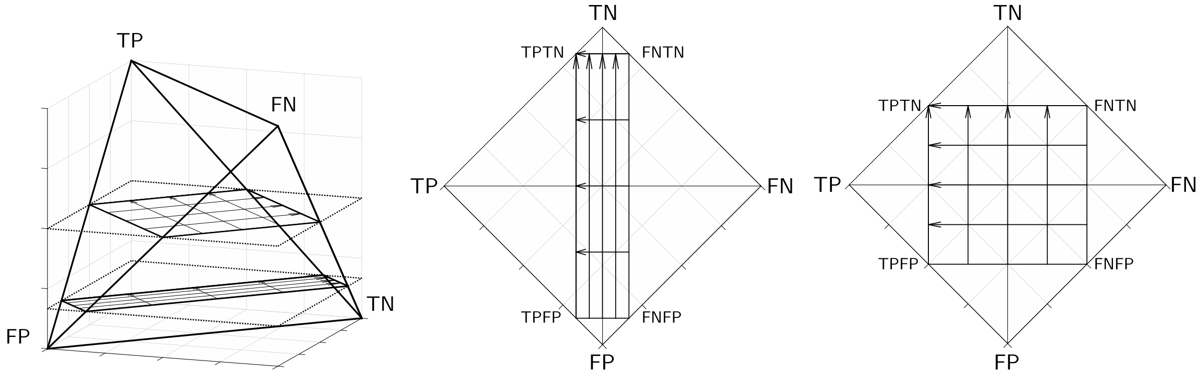
Fig. 8. Illustration of properties $\overline{\mathsf{FN}}_{min}$ and $\overline{\mathsf{FP}}_{min}$

Properties $TP_{\nearrow}$ and $TN_{\nearrow}$ require that growing TP and TN values (i.e., improving correct predictions of the classifier) should coincide with a weakly monotonic growth of the measure's value (Fig. 9). As far as $TP_{\nearrow}$ is concerned, observe that the greater TP is in the confusion matrix, the closer we move from side $\overline{\mathsf{FN}}$ to side $\overline{\mathsf{TP}}$ in the rectangular cross-section, which translates directly to increased recognition of the positive class. Naturally, it would be counter-intuitive if such increased recognition resulted in decreasing values of a classification measure. Thus, its weakly monotonic growth is expected. As opposed to requirements $\overline{\mathsf{FN}}_{min}$ and $\overline{\mathsf{FP}}_{min}$, which concern merely the borders of the cross-section, $TP_{\nearrow}$ concerns the entirety of the cross-section. These properties resolve to the following conditions:

$$TP_{\nearrow} : \text{if } TP_1 \geq TP_2, \text{ then } f(\left[\begin{smallmatrix} TP_1 & FN_1 \\ FP & TN \end{smallmatrix}\right]) \geq f(\left[\begin{smallmatrix} TP_2 & FN_2 \\ FP & TN \end{smallmatrix}\right]),$$

$$TN_{\nearrow} : \text{if } TN_1 \geq TN_2, \text{ then } f(\left[\begin{smallmatrix} TP & FN \\ FP_1 & TN_1 \end{smallmatrix}\right]) \geq f(\left[\begin{smallmatrix} TP & FN \\ FP_2 & TN_2 \end{smallmatrix}\right]).$$

Properties $\overline{\mathsf{TN}}_{\neq max}$ and $\overline{\mathsf{TP}}_{\neq max}$ tackle the problem of maximal values of the measure. Observe that in a two-class problem, the full recognition of just one class (only positive or only negative), which can be achieved trivially, should not render the highest value of the measure. Only the full recognition of both classes should be rewarded with the maximum, as stated by TPTN_{max}. Thus, properties $\overline{\mathsf{TN}}_{\neq max}$ and $\overline{\mathsf{TP}}_{\neq max}$ require that the measure's values on sides $\overline{\mathsf{TN}}$ and $\overline{\mathsf{TP}}$ should be less than maximal, except for



(a) External view **(b)** $P/n = 1/6$ **(c)** $P/n = 1/2$

Fig. 9. Illustration of properties $TP_{\nearrow}$ and $TN_{\nearrow}$

the very vertex TPTN. If a classification measure fulfils this property, a simple majority or minority stub (always predicting the same class) will never be mistaken for the best possible classifier. This boils down to:

$$\overline{\mathsf{TN}}_{\neq max},\ \overline{\mathsf{TP}}_{\neq max} : \text{if } FN + FP > 0, \text{ then } f(\left[\begin{smallmatrix} TP & FN \\ FP & TN \end{smallmatrix}\right]) < max.$$

Property ACE reveals the class bias resulting from *asymmetric class evaluation*, typical for class imbalance problems where the minority class is treated as the more important one. It is introduced to guarantee that full recognition of only the negative class is never rewarded with a higher value than the full recognition of only the positive one (assuming the respective other class is recognized to the same degree). In particular, since the recognition of the positive class is of high importance, the middle point of side $\overline{\mathsf{TP}}$ (i.e. when the whole positive and half of the negative class is recognized) should not be assessed with a lower value than the middle point of side $\overline{\mathsf{TN}}$ (i.e. when the whole negative and half of the positive class is recognized). The same expectations can be formulated for all other pairs of corresponding points on sides $\overline{\mathsf{TP}}$ and $\overline{\mathsf{TN}}$ (three of which are depicted in Fig. 10). In terms of the entries of the confusion matrix:

$$ACE : f(\left[\begin{smallmatrix} P & 0 \\ \gamma N & (1-\gamma)N \end{smallmatrix}\right]) \geq f(\left[\begin{smallmatrix} (1-\gamma)P & \gamma P \\ 0 & N \end{smallmatrix}\right]),$$

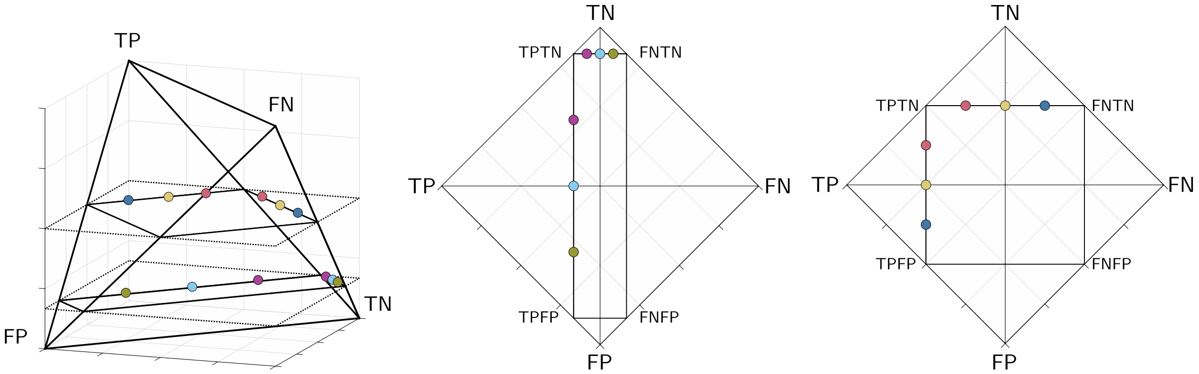
where $\gamma \in [0, 1]$ (in Fig. 10 γ takes on values $1/4$, $2/4$ and $3/4$). Notice that the weak nature of the property is implied by the fact that it does not specify by how much the full recognition of the positive class should be favoured over the full recognition of the negative class. On the other hand, the unsatisfied ACE reveals instantly that the measure favours the negative over the positive.

Much the same, property ACH deals with aspects of *asymmetric class handling*. It tests if the classes can be exchanged without influencing the measure's behaviour. This could be relevant in some sub-categories of class imbalance problems, for instance in data streams plagued by concept drift [6]. In such dynamic environments, the percentage of the positive class may increase to make it actually (albeit temporarily) the majority class [42]. Expressed with the confusion matrix:

$$ACH : f(\left[\begin{smallmatrix} TP & FN \\ FP & TN \end{smallmatrix}\right]) = f(\left[\begin{smallmatrix} TN & FP \\ FN & TP \end{smallmatrix}\right]).$$

Finally, property *UnDefs* pinpoints the existence and the location of undefined values ($+\infty$ whenever a positive value is divided by 0, $-\infty$ whenever a negative value is divided by 0, and NaN whenever 0 is divided by 0). As such, it highlights potential numerical pitfalls that can arise when calculating the measure values. While occurring fairly seldom with real life data, such undefined values are needed to fully characterize and thoroughly compare the considered measures.

As the introduced properties aim to support researchers in analysing and comparing measures, we will briefly discuss some implications of the properties for exemplary practical situations of solving class imbalanced tasks. Although one cannot expect a single best measure for all possible cases, we will provide some guidelines for particular sub-cases.



(a) External view **(b)** $P/n = 1/6$ **(c)** $P/n = 1/2$

Fig. 10. Illustration of property *ACE*

Assume that an imbalanced dataset contains two classes, one of which is significantly more important to the user. For instance, it may be the case of medical diagnosis problems where commonly false negatives (ill patients classified as healthy) are more risky than false positives (healthy patients classified as ill). Similar scenarios occur in many other domains such as financial analysis, fraud detection, technical diagnostics [22]. In all such situations, the choice of a good measure should mainly be influenced by the possession of property *ACE*, which represents the asymmetry between classes. Additionally, properties that refer to taking maximal or minimal measure values, i.e. TPTN_{max}, $\overline{\mathsf{TN}}_{\neq max}$ or $\overline{\mathsf{TP}}_{\neq max}$ should not be overlooked.

On the other hand, the *ACH* property should definitely be considered when choosing a measure for classification of evolving data streams when the classes are imbalanced and change their cardinalities with time [6]. When one class is heavily under-represented, many classifiers tend to be biased towards the majority class, thus the symmetry handling property should be analysed.

Moreover, particular attention should be paid to the *UnDefs* property when the measure is applied to a dataset jeopardised by extreme class under-representation, resulting in the total absence of at least one class. Should one encounter such danger in the application at hand, measures featuring many occurrences of undefined values ought to possibly be avoided. Situations of temporal class disappearance may, for instance, occur while processing data from sensors.

Finally, slightly different considerations concern situations when the measure controls classifier tuning. During classifier tuning, different sets of parameters are systematically tested to find the set which maximizes the performance measure. These procedures require proper variability of the classification measure, thus measures possessing monotonicity properties $TP_{\nearrow}$ and $TN_{\nearrow}$ are strongly recommended.

The impact of the proposed ten properties on comparisons of various classification measures will be studied in the next sections.

5. Analysis of Classification Measures with Respect to their Properties

Having presented the visualization technique in Section 3 and having defined ten properties in Section 4, now we use the proposed tools to analyse 22 classification measures. The results of this analysis should provide researchers with informed means of measure comparison and comprehensive knowledge on their behaviour. The analysed measures, widely discussed in the literature in the context of binary imbalanced data, have been chosen for their diversity and popularity. Moreover, basic measures for tasks without class imbalance, such as classification accuracy, have been added to complement the measure comparison.

All of the analysed measures are defined using entries of a two class confusion matrix (see Section 2). The vast majority of these measures are calculated solely on the basis of the confusion matrix entries and shall be referred to as the non-parametric ones (e.g. *sensitivity* and *specificity*). Contrastingly, complex measures that employ additional parameters will be called parametric (e.g. F_β). The detailed definitions of all 22 considered measures are available in the supplementary materials[2].

Table 2 summarizes the results of the verification of the ten properties for each of the examined measures. The properties clearly differentiate the analysed measures. Below we discuss some of these differences, providing interesting, and sometimes unexpected observations that should aid researchers in the process of measure selection for their tasks.

Table 2: Properties of selected classification measures; *: contains NaN (undefined value); †: NaN side, ^{s}: strong monotonicity

Measure	TPTN_{max}	$\overline{\mathsf{FN}}_{min}$	$\overline{\mathsf{FP}}_{min}$	$TP_{\nearrow}$	$TN_{\nearrow}$	$\overline{\mathsf{TN}}_{\neq max}$	$\overline{\mathsf{TP}}_{\neq max}$	ACE	ACH	$UnDefs$
Accuracy	$\checkmark$	$\times$	$\times$	$\checkmark^{s}$	$\checkmark^{s}$	$\checkmark$	$\checkmark$	$\times$	$\checkmark$	none
Area Under Lift	$\times$	$\times$	$\times$	$\checkmark^{s}$	$\checkmark^{s}$	$\checkmark$	$\checkmark$	$\checkmark$	$\times$	TN–FP; TP–FN
Balanced accuracy	$\checkmark$	$\times$	$\times$	$\checkmark^{s}$	$\checkmark^{s}$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	TN–FP; TP–FN
F_1*-score*	$\checkmark$	$\times^{\dagger}$	$\times$	$\checkmark^{s*}$	$\checkmark^{s*}$	$\checkmark^{*}$	$\checkmark$	$\times$	$\times$	△FP–FN–TN
False negative rate	$\times$	$\times$	$\times$	$\times$	$\checkmark$	$\times$	$\checkmark$	$\times$	$\times$	TN–FP
False positive rate	$\times$	$\times$	$\times$	$\checkmark$	$\times$	$\checkmark$	$\times$	$\checkmark$	$\times$	TP–FN
F_β, $\beta \in [0,\infty)$	$\checkmark$	$\times^{\dagger}$	$\times$	$\checkmark^{s*}$	$\checkmark^{s*}$	$\checkmark^{*}$	$\checkmark$	$\times$	$\times$	△FP–FN–TN
G-mean	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	TN–FP; TP–FN
IBA_α(*Accuracy*), $\alpha \in (0,\infty)$	$\times$	$\times$	$\times$	$\checkmark^{s}$	$\times$	$\checkmark$	$\times$	$\times$	$\times$	TN–FP; TP–FN
IBA_α(F_1*-score*), $\alpha \in (0,\infty)$	$\times$	$\times^{\dagger}$	$\times$	$\times$	$\times$	$\checkmark^{*}$	$\times$	$\times$	$\times$	△FP–FN–TN; TP–FN
IBA_α(*G-mean*), $\alpha \in (0,\infty)$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$	$\times$	$\checkmark$	$\times$	$\checkmark$	$\times$	TN–FP; TP–FN
$IBA_\alpha(F_\beta)$, $\alpha,\beta \in (0,\infty)$	$\times$	$\times^{\dagger}$	$\times$	$\times$	$\times$	$\checkmark^{*}$	$\times$	$\times$	$\times$	△FP–FN–TN; TP–FN
Jaccard coefficient	$\checkmark$	$\checkmark$	$\times$	$\checkmark^{s}$	$\checkmark$	$\checkmark$	$\checkmark$	$\times$	$\times$	TN
Kappa	$\checkmark$	$\times$	$\times$	$\checkmark^{s}$	$\checkmark^{s}$	$\checkmark$	$\checkmark$	$\times$	$\times$	TN; TP
Log odds-ratio	$\checkmark$	$\checkmark^{*}$	$\checkmark^{*}$	$\checkmark^{*}$	$\checkmark^{*}$	$\times$	$\times$	$\times^{\dagger}$	$\checkmark$	TN–FN; TN–FP; TP–FN; TP–FP
MCC	$\checkmark$	$\times$	$\times$	$\checkmark^{s*}$	$\checkmark^{s*}$	$\checkmark^{*}$	$\checkmark^{*}$	$\times$	$\checkmark$	TN–FN; TN–FP; TP–FN; TP–FP
Neg. predictive value	$\checkmark$	$\times$	$\checkmark^{*}$	$\checkmark^{*}$	$\checkmark^{*}$	$\checkmark$	$\times$	$\checkmark$	$\times$	TP–FP
OP	$\checkmark$	$\times$	$\times$	$\times$	$\times$	$\checkmark$	$\checkmark$	$\times$	$\checkmark$	TN–FP; FP–FN; TP–FN
Pointwise AUC-ROC	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	TN–FP; TP–FN
Precision	$\checkmark$	$\checkmark^{*}$	$\times$	$\checkmark^{*}$	$\checkmark^{*}$	$\times$	$\checkmark$	$\times$	$\checkmark$	TN–FN
Recall	$\checkmark$	$\checkmark$	$\times$	$\checkmark^{s}$	$\checkmark$	$\checkmark$	$\times$	$\checkmark$	$\times$	TN–FP
Specificity	$\checkmark$	$\times$	$\checkmark$	$\checkmark$	$\checkmark^{s}$	$\times$	$\checkmark$	$\times$	$\times$	TP–FN

Let us start with the TPTN_{max} property (Table 2). Recall that it ensures obtaining the best measure value in case of perfect predictions of both classes. Interestingly however, some measures designed for imbalanced data, such as e.g. IBA_α(*G-mean*), IBA_α(*Accuracy*), IBA_α(F_1*-score*) do not satisfy this quite natural requirement. A more detailed explanation of *IBA* parametrization, including the cause of the underlined deterioration of the maximum value, is provided in the next section.

Surprisingly, the next properties $\overline{\mathsf{FN}}_{min}$ and $\overline{\mathsf{FP}}_{min}$ are also not met by many measures considered in studies on imbalanced data. As justified in Section 4, classifiers failing to recognize one of the classes are not acceptable in many applications. Thus, measures that obtain non-minimal values when one of the classes is completely misclassified (e.g. F_1*-score*, *OP*, *MCC*) should be avoided or used with greater care.

Looking at the monotonic properties $TP_{\nearrow}$ and $TN_{\nearrow}$ one can notice that they are valid for nearly all measures. Only very few measures like *OP* fail to satisfy the weakly monotonic growth of values when the

[2] https://dabrze.shinyapps.io/Tetrahedron/

correct predictions of the classifier are increasing. More detailed investigations concerning the monotonic changes for particular measures and particular class proportions can be conducted using the proposed visualization technique as exemplified in the next section.

Moving on to properties $\overline{\mathsf{TN}}_{\neq max}$ and $\overline{\mathsf{TP}}_{\neq max}$ one can notice that *Log odds-ratio* should be used with much care as it does not satisfy any of them. Consequently, it can obtain maximal values even though the full recognition of both classes has not been achieved. Table 2 also indicates other measures that do not meet either $\overline{\mathsf{TN}}_{\neq max}$ or $\overline{\mathsf{TP}}_{\neq max}$ (e.g. *Precision*, *Recall*) advising researchers to be particularly cautious while choosing them for their practical learning tasks.

The ACE property, concerning asymmetric approach to classes (in favour of the minority class) is another factor strongly differentiating all the measures. Although improving recognition of the minority class, even at the cost of slight worsening of the recognition of the majority examples is a common goal in most classification methods for imbalanced data, several measures often exploited in the literature (even as popular as F_1*-score* or *Kappa* statistic) do not satisfy this property. As a result, these measures can actually favour the negative class, as it is easier to increase their value by improving recognition of the majority class than that of the minority class.

Interestingly, the other property related to asymmetry, i.e. ACH, is rarely satisfied by the selected measures. Consequently, the behaviour of many measures changes when the class proportions are switched. Researchers working with highly dynamic data should, therefore, focus on measures satisfying ACH.

Finally, let us discuss the existence and location of undefined values (*UnDefs*). Undefined measure values, usually resulting from division by zero, are commonly neglected. Nevertheless, they may well occur with imbalanced data, for example during cross-validation procedures when one of the two classes happens to be unrepresented in the validation set. The problem becomes aggravated for multi-class imbalanced data when a measure is macro-averaged for all classes, since the resulting average becomes undefined if at least one of the averaged values is undefined. An interesting observation is that, except for *accuracy*, all of the considered measures contain undefined values. In particular, the *Kappa* statistic is undefined when there exist only positive or only negative examples in the dataset and none of them is misclassified, which translates directly to vertex TP and vertex TN in the tetrahedron. Even worse, *balanced accuracy* is undefined when there are only positive or only negative examples in the dataset, which translates to edges TP–FN and TN–FP. Worst of all, F_1*-score* (as well as its generalizations) exhibits undefined values in face △FP–FN–TN, which occurs when all positive examples are misclassified (even when both classes are represented).

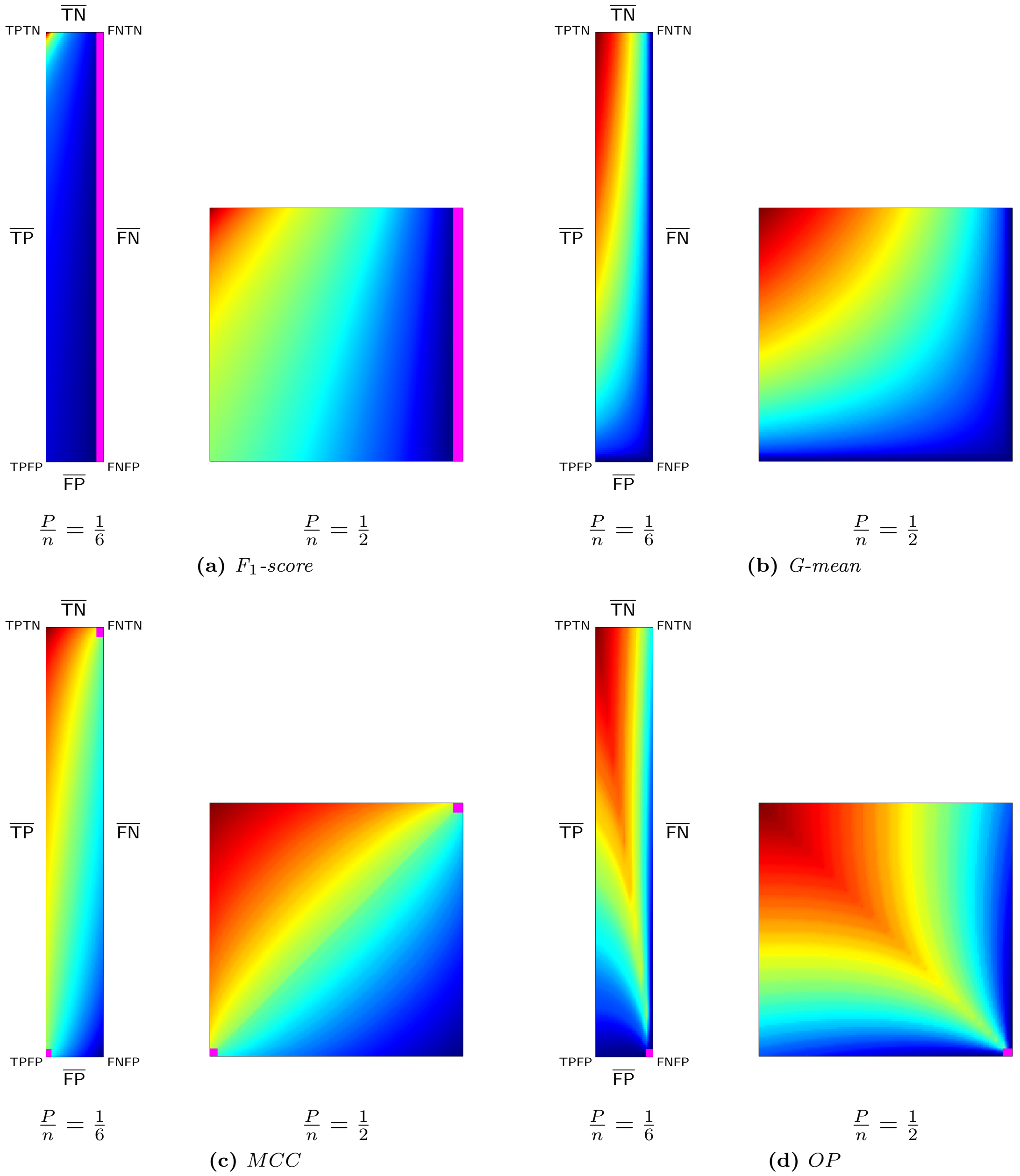
The analysis presented in this section indicates the importance of studying measure properties, particularly when comparing various measures proposed in the literature. The discussed results extend current studies on classification performance measures for imbalanced data, and show that the devised properties can aid machine learning experts during measure selection.

6. Visual-based Analysis of Selected Measures

This section includes case studies of using our visualization tool to examine more thoroughly differences between few selected measures in complete ranges of their values. Referring to properties presented in Table 2, we will demonstrate additional details, such as local changes of measures values, monotonicity of these changes, and locations of atypical values. The first subsection will focus on non-parametric measures, while the second one will concern more complex measures with compound parametric definitions. We will show how the visual inspection can support examining the impact of the internal parametrization in the F_β measure and the external parametrization in IBA_α(*G-mean*). Finally, our observations will lead us to analytical derivations of threshold bounds for parameter values of these measures.

6.1. Non-parametric Measures

Let us now conduct a more detailed visual analysis of four measures highlighted in Section 2: F_1*-score*, *G-mean*, *Mathews Correlation Coefficient* (*MCC*), and *Optimized precision* (*OP*). Due to the page limit, in this paper we present only two cross-sections produced for $P/n = 1/6$ and $P/n = 1/2$. However, other cross-sections of the tetrahedron, including cross-sections produced for higher levels of class imbalance, can be viewed in the online visualization tool.



(a) F_1*-score*

(b) *G-mean*

(c) *MCC*

(d) *OP*

Fig. 11. Cross-sections of selected measures for $P/n = 1/6$ and $P/n = 1/2$

The visual analysis of F_1*-score* (Fig. 11a) shows that the growth (although monotonic) of the measure along side $\overline{\text{TP}}$ is very slow and does not fulfil the ACE property when the data are imbalanced. To illustrate this, consider Fig. 11a (left), which corresponds to class imbalance, and a point located in the middle of $\overline{\text{TP}}$. The value there is much lower than the corresponding point on side $\overline{\text{TN}}$. Taking into account the fact that the middle point of $\overline{\text{TP}}$ corresponds to full recognition of the positive class and 50% recognition of the negative class, this shows that with class imbalance high values corresponding to full recognition of the positive class are harder to obtain. Expressed in terms of values in the confusion matrix:

$$f(\begin{bmatrix} P & 0 \\ \gamma N & (1-\gamma)N \end{bmatrix}) < f(\begin{bmatrix} (1-\gamma)P & \gamma P \\ 0 & N \end{bmatrix}),$$

where $\gamma = 1/2$. Notice that while F_1-*score* fulfils ACE for $P/n = 1/2$ (Fig. 11a (right)), it does not for the above mentioned $P/n = 1/6$ (Fig. 11a (left)), which means that the property is not satisfied in general (i.e. throughout the tetrahedron). Evidently, the property cannot be verified using only one selected cross-section. As may be observed using the online tool (in particular, by animating P/n from $1/2$ down to 0), this flawed feature of the measure aggravates for increasing class imbalance (i.e. when P/n drops). This may be quite surprising as F_1-*score* is often brought out in the literature as especially suited for the positive (minority) class. Generalizations of F_1-*score*, which could overcome some of these limitations, will be discussed in subsection 6.2 devoted to parametric measures.

On the other hand, the visual-based analysis of *G-mean* (Fig. 11b) reveals that this measure satisfies the devised properties. In particular, it satisfies some important properties not fulfilled by F_1-*score*, *MCC* and *OP*. First, as opposed to the other three measures, *G-mean* takes minimal values on whole sides $\overline{\mathsf{FN}}$ and $\overline{\mathsf{FP}}$. Additionally, it enjoys the ACE property, which makes the measure especially useful in the contexts of imbalanced data: for any two corresponding points on sides $\overline{\mathsf{TP}}$ and $\overline{\mathsf{TN}}$, the value on side $\overline{\mathsf{TP}}$ happens to be equal (and thus not smaller) to that on $\overline{\mathsf{TN}}$. This means that for any γ:

$$f(\left[\begin{smallmatrix} P & 0 \\ \gamma N & (1-\gamma)N \end{smallmatrix}\right]) = f(\left[\begin{smallmatrix} (1-\gamma)P & \gamma P \\ 0 & TN \end{smallmatrix}\right]).$$

Analysing *MCC* (Fig. 11c), one can observe that its values on sides $\overline{\mathsf{FP}}$ and $\overline{\mathsf{FN}}$ are not minimal, which violates properties $\overline{\mathsf{FN}}_{min}$ and $\overline{\mathsf{FP}}_{min}$. Moreover, comparing cross-sections for $P/n = 1/2$ and $P/n = 1/6$, one can observe that small values are harder to obtain with the increase of class imbalance. Furthermore, similarly to F_1-*score*, *MCC* does not satisfy the previously discussed property ACE. Even though for balanced classes (Fig. 11c (right)) the corresponding points in $\overline{\mathsf{TP}}$ and $\overline{\mathsf{TN}}$ feature equal values, this deteriorates with growing disproportion between classes (Fig. 11c (left)). In other words, for imbalanced data it is easier to obtain undue high values by recognizing the negative class, which contradicts typical expectations.

Finally, let us consider measure OP (Fig. 11d). The visual-based analysis reveals that OP is the only of the considered measures that does not satisfy properties $TP_{\nearrow}$ and $TN_{\nearrow}$. Observe that traversing the cross-sections horizontally right-to-left or vertically bottom-up (thus increasing the recognition of one of the classes while keeping the recognition of the second one constant) the values of the measure first increase and then decrease. In fact, the visual analysis discloses that the measure is designed to increase its values monotonically only when the recognition of both classes increases. Undeniably, the increase of the recognition of both classes at the same time is highly desirable and should imply increasing measure values, however, the observed behaviour of OP in (acceptable) cases when the classifier increases the recognition of one class while keeping the recognition of the other constant is rather surprising and counter-intuitive. This discussion shows the practical importance of $TP_{\nearrow}$ and $TN_{\nearrow}$ in the context of controlling the classifier tuning procedure. The information on whether a measure possesses these properties or not is thus indispensable in applications where classification parameters are optimized in a process driven by the selected measure.

6.2. Parametric Measures

Up to now we have considered measures defined directly as functions of the four entries of the confusion matrix. Besides such unparametrized formulas one can distinguish complex measures with parameters that control the trade off between predictions referring to different classes. For instance, F_β is the generalized form of F_1-*score*, where the β parameter controls the aggregation of *precision* and *recall*. As such control is much desired, other external parametrization procedures have also been developed to modify measures for imbalanced data. A representative method of this approach is the *Index of Balanced Accuracy* (IBA_α) [17, 18, 19], which is capable of adapting any classification measure to the imbalanced domain by introducing an α parameter that controls the amount by which the original measure is actually modified.

The above approaches allow us to focus on two parametrization types:

- internal parametrization (e.g. F_β),
- external parametrization (e.g. $IBA_\alpha(G\text{-}mean)$).

Observe that measure parametrization actually increases the number of available degrees of freedom, making the inherently complex analyses of such measures even more challenging than in case of basic non-parametrized measures. The principal questions in this context are: how are the particular parameter values to be established? And further, what are their applicability ranges?

To answer these questions, let us now conduct a more detailed visual analysis of measures, representing both types of parametrization: internal (F_β) and external ($IBA_\alpha(G\text{-}mean)$). Consistently, we present the impact of the parametrization using cross-sections produced for $P/n = 1/6$ and $P/n = 1/2$, whereas other cross-sections and entire tetrahedrons can be viewed in our online visualization tool.

6.2.1. Internal parametrization: F_β

While F_1*-score* is a regular harmonic mean of *precision* and *recall*, F_β originated as a weighed version of this mean. In F_β, λ and $1-\lambda$ act as non-negative ($0 \le \lambda \le 1$) weights of *precision* and *recall*, respectively. This means that λ may be chosen to produce any convex combination of $1/$*precision* and $1/$*recall* to be actually used in the mean. Let p denote *precision* and r denote *recall*, the weighted harmonic mean of p and r is: $(\frac{\lambda\frac{1}{p}+(1-\lambda)\frac{1}{r}}{\lambda+(1-\lambda)})^{-1} = \frac{\lambda+(1-\lambda)}{\lambda\frac{1}{p}+(1-\lambda)\frac{1}{r}} = \frac{1}{\lambda\frac{1}{p}+(1-\lambda)\frac{1}{r}} = \frac{1}{\lambda\frac{r}{pr}+(1-\lambda)\frac{p}{pr}} = \frac{pr}{\lambda r+(1-\lambda)p} = \frac{\frac{1}{\lambda}pr}{r+\frac{1-\lambda}{\lambda}p}$ (from now on: $\lambda > 0$).

After setting[3] $\beta = \frac{1-\lambda}{\lambda}$, one gets $\frac{1}{\lambda} = \beta + 1$, which finally produces: $F_\beta = \frac{(\beta+1)pr}{\beta p+r}$. Notice that in this scheme:

- $\lambda \to 0$ corresponds to $\beta \to \infty$ (emphasis on *precision*),
- $\lambda = 0.5$ corresponds to $\beta = 1.0$ (equal emphasis),
- $\lambda \to 1$ corresponds to $\beta \to 0$ (emphasis on *recall*).

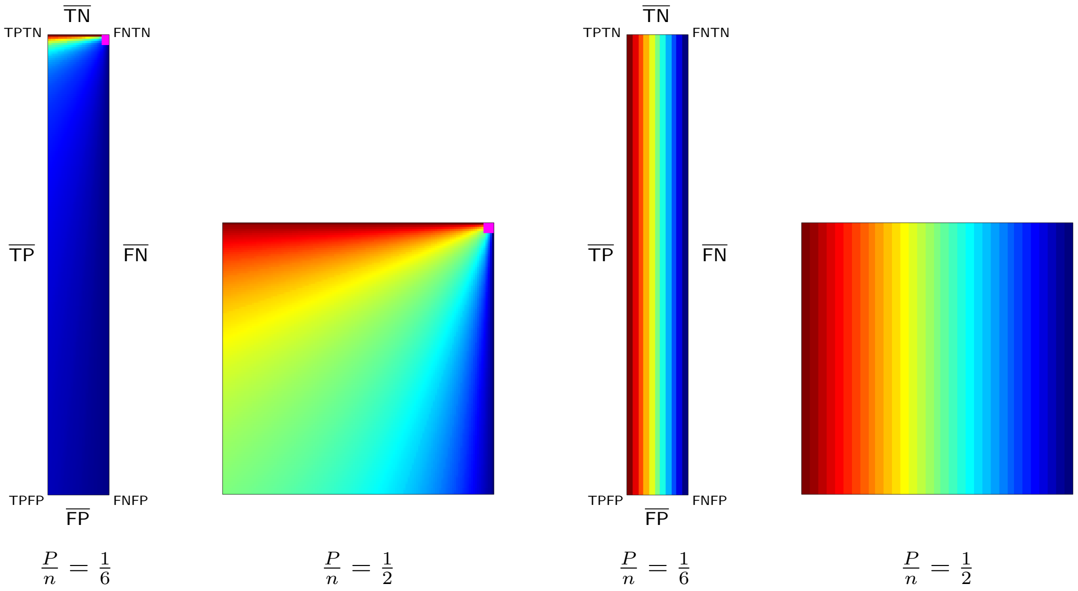
Of course, for $\beta = 1.0$, measure F$_\beta$ becomes $\frac{(1+1)pr}{1\cdot p+r} = 2\frac{pr}{p+r} = F_1$*-score*, which is thus the regular (unweighed) harmonic mean of *precision* and *recall*.

However, both precision and recall could be aggregated in different ways [25]. The harmonic mean, used in this context happens to be the most conservative of the three popular Pythagorean means: arithmetic (A), geometric (G) and harmonic (H), as they satisfy $A \ge G \ge H$, but it is also easy to visualize the two others in this role. To what extent and in which regions of the domain these three different means of *precision* and *recall* actually diverge from one another may be observed in Fig. 12 and 13, where both *precision* and *recall* as well as their three means (arithmetic: $A(p,r)$, geometric: $G(p,r)$ and harmonic: $H(p,r)$) are shown. This visualization illustrates well the concave isolines of $A(p,r)$ and $G(p,r)$, which means that they obtain excessively high values for increasingly divergent recognition of classes, making $H(p,r)$ the best choice out of three in this respect.

Deciding on the mean, however, is not enough, as the remaining problem regards changes in the measure's behaviour across the differing minority ratio P/n. Unfortunately, for the harmonic mean as well as for the other two means, the measure's values gradually shift away from the positive class as P/n decreases, making all three (regular) means of *precision* and *recall* (and thus the F_1*-score* in particular) less and less suited for imbalanced data. This is why the weighed means, in particular F_β, may actually turn out to be more useful.

The arising question regards the appropriate value of β. Clearly, the desired bias towards the positive class requires $\beta > 1$, which corresponds to applying more weight to *recall*. The visual solution to this problem is provided in Fig. 14, which shows cross-section visualizations of F_β for three values of $\beta \in \{1, 3, 5\}$. The range of these values has been inspired by the accessible data, in this case the class ratios considered in previous sections: $P/n = 1/6$ and $P/n = 1/2$. These values may be assumed to directly express the $[0,1]$-based weights of *precision* and *recall*, i.e. $\lambda = 1/6$ and $\lambda = 1/2$, which translate to $\beta = 5$ and $\beta = 1$, respectively.

[3] Some authors set $\beta = \sqrt{\frac{1-\lambda}{\lambda}}$ instead, resulting in $\frac{1-\lambda}{\lambda} = \beta^2$, which allows for some further interpretation of such β [38]; not to be pursued in this paper.



(a) *precision* **(b)** *recall*

Fig. 12. Cross-sections of *precision* and *recall*

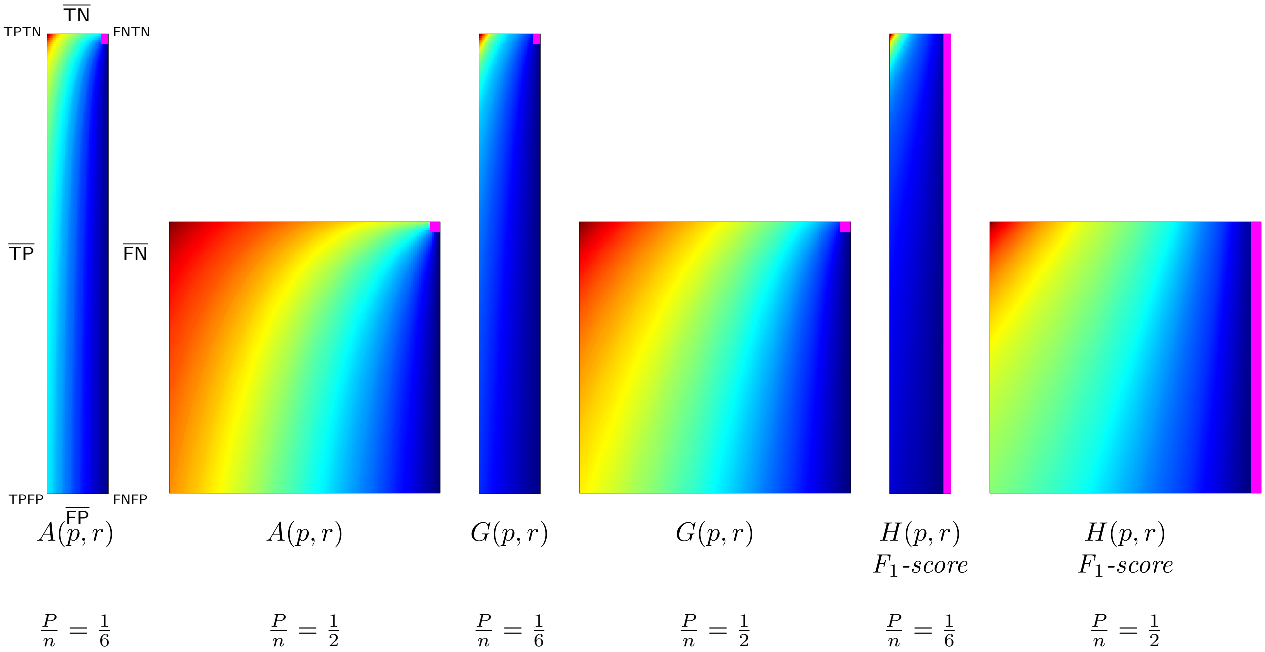


Fig. 13. Cross-sections of different means of *precision* and *recall*

Despite the fact that all ten of the earlier discussed properties of F_1-*score* and F_β (for $\beta \in [0, \infty)$) are identical (Table 2), the increasing difference between F_1-*score* and F_β resulting from the changing values of β is clearly visible in Fig. 14, revealing the truly multi-dimensional complexity of the measures' domains.

A slightly closer explanation may only be due for ACE property, as F_β's particular visualization for $\beta = 5$ in Fig. 14 suggests that the measure satisfies the requirements of ACE (its values on the $\overline{\text{TP}}$ side are not lower than their counterparts on the $\overline{\text{TN}}$ side for both $P/n = 1/6$ and $P/n = 1/2$), whereas Table 2

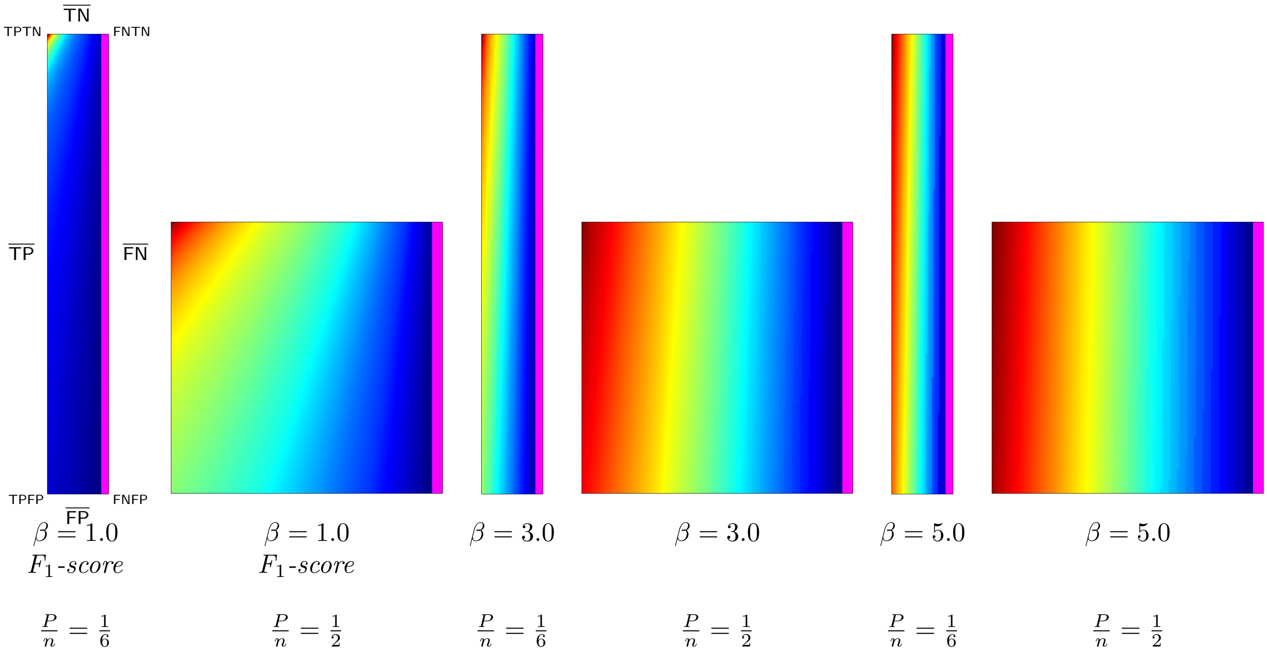


Fig. 14. Cross-sections of F_β

states that ACE is not met by F_β. This is because the ten proposed properties are of general character, i.e. they concern the whole tetrahedron, which means that they must be satisfied in cross-sections corresponding to all feasible class proportions. In case of $F_{\beta=5}$, for some class ratios that are lower than those considered in the presented visualizations, e.g. for $P/n = 1/10$ (easily reproducible in the online visualization tool), the ACE conditions are actually not satisfied, thus justifying the contents of Table 2.

Nevertheless, for cases when the class ratio is known or predictable, the visualizations are of utmost practical value. In the discussed situation, the visual-based analysis may suggest non-trivial values of β for which F_β certainly satisfies selected properties, in this case the conditions of ACE for a particular P/n. A thorough analysis of cross-sections clearly suggested the existence of a particular dependency between β and the class proportion, which influences the ACE property. This observation inspired us to derive analytically the borderline value of β that ensures that ACE is met by F_β.

Proposition 1. *F_β satisfies ACE property for $\beta \geq N/P$ (for proof see the Appendix).*

Practically this means that the user must bear in mind the class proportions and may use it to make F_β satisfy the ACE property, if needed.

6.2.2. External parametrization: IBA_α(G-mean)

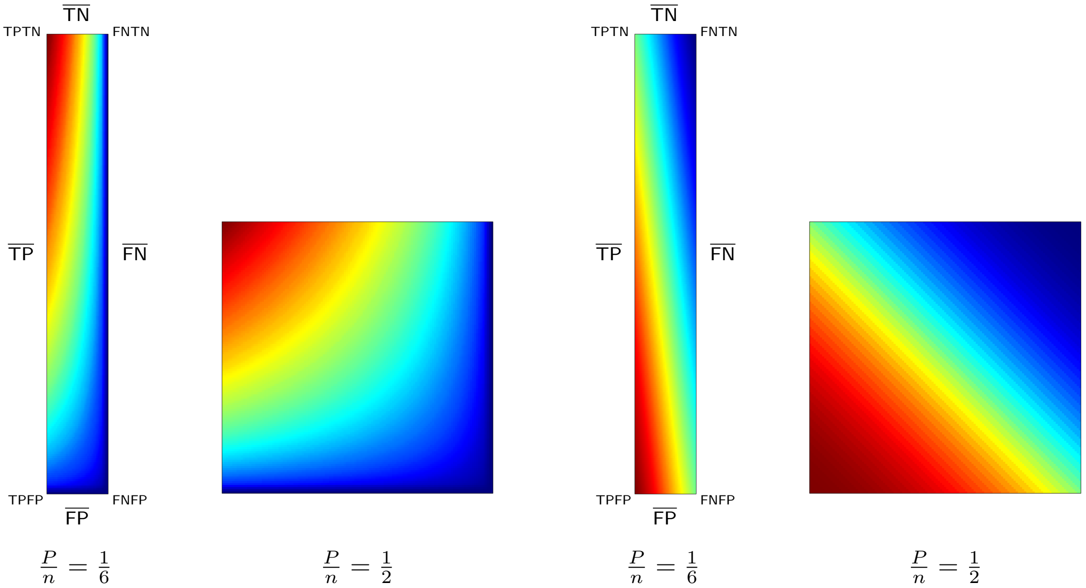
Applying any external parametrization, e.g. the IBA_α scheme [17, 18, 19], to different measures evokes several issues, first of all related to establishing the values of required parameters. Our visualization provides a very practical solution to these issues, as shall be demonstrated in this section.

Given a classifier performance measure M, a parameter $\alpha \geq 0$ and a tentative measure $Dom = sensitivity - specificity$, the formula [17]:

$$IBA_\alpha(M) = (1 + \alpha Dom)M$$

defines the parametrization of M, in which this measure is multiplicatively combined with $(1 + \alpha Dom)$. Of course, for $\alpha = 0$: $IBA_\alpha(M) = M$. Simultaneously, when $Dom \in [-1, +1]$ and $\alpha \leq 1$ then $1 + \alpha Dom \geq 0$, which, together with $M \geq 0$, implies $IBA_\alpha(M) \geq 0$.

The scheme has been conceived to increase the measure orientation towards the positive class, which is preferred in most imbalanced classification problems. Notice, however, that neither *Dom* is a classic classifier performance measure (as its domain includes negative values), nor is $IBA_\alpha(M)$ a simple convex combination of *Dom* and M. This renders strictly analytical (without any visualization tool) analysis of $IBA_\alpha(M)$ very hard, especially for larger values of α. In result, while the general goal of reorienting the measure towards the positive class is certainly achieved by IBA_α, it is not instantly clear how this reorientation is practically manifested. In particular, one might be interested in identifying whether measure M subjected to $IBA_\alpha(M)$ satisfies any of the postulated properties, or not (and, if it does, which ones and for what ranges of α).



(a) *G-mean* **(b)** *Dom*

Fig. 15. Cross-sections of *G-mean* and *Dom*

Below, we visualize and analyse *G-mean* externally parametrized according to IBA_α for $\alpha \in \{0, 0.5, 1\}$. The combination IBA_α(*G-mean*) was particularly recommended and analytically studied for the aforementioned α values by García et al. [17]. Tracing the influence of *Dom* on *G-mean* within the IBA_α approach may well be started with the visualization of the components of the parametrization procedure (Fig. 15). Clearly, for $\alpha \to 0$, IBA_α(*G-mean*)$\to$ *G-mean*, so only $\alpha > 0$ exerts any influence on the result. Notice that *Dom* features a rather unexpected growth towards vertex TPFP, implying the specific behaviour of IBA_α(*G-mean*), see Fig. 16. Because the combination is multiplicative, the values of *G-mean* are being 'amplified' by the corresponding values of $(1 + \alpha Dom)$, in particular: increased for $(1 + \alpha Dom) > 1$, and decreased for $(1 + \alpha Dom) < 1$.

As stated in Table 2, *G-mean* satisfies all of the proposed properties. The important question is how the application of external parametrization to the measure influences its properties, e.g. the ACE property. Unsurprisingly, IBA_α(*G-mean*) may be proven to satisfy the ACE property for all assumed values of α. Notably, this comes with a cost, as this parametrization of *G-mean* is not equally stable with respect to all other properties.

Proposition 2. *IBA_α(G-mean) satisfies ACE property for $\alpha \geq 0$ (for proof see the Appendix).*

Practically, this means that the IBA_α(*G-mean*) does not depart from the original *G-mean* in terms of ACE. However, a thorough visual-based analysis of the impact of the α parameter on satisfying $TN_{\nearrow}$ by IBA_α(*G-mean*) suggested a border-line value of α. Inspired thereby, we derived the exact value analytically.

Proposition 3. *IBA_α(G-mean) satisfies $TN_{\nearrow}$ property for $\alpha \leq 1/3$ (for proof see the Appendix).*

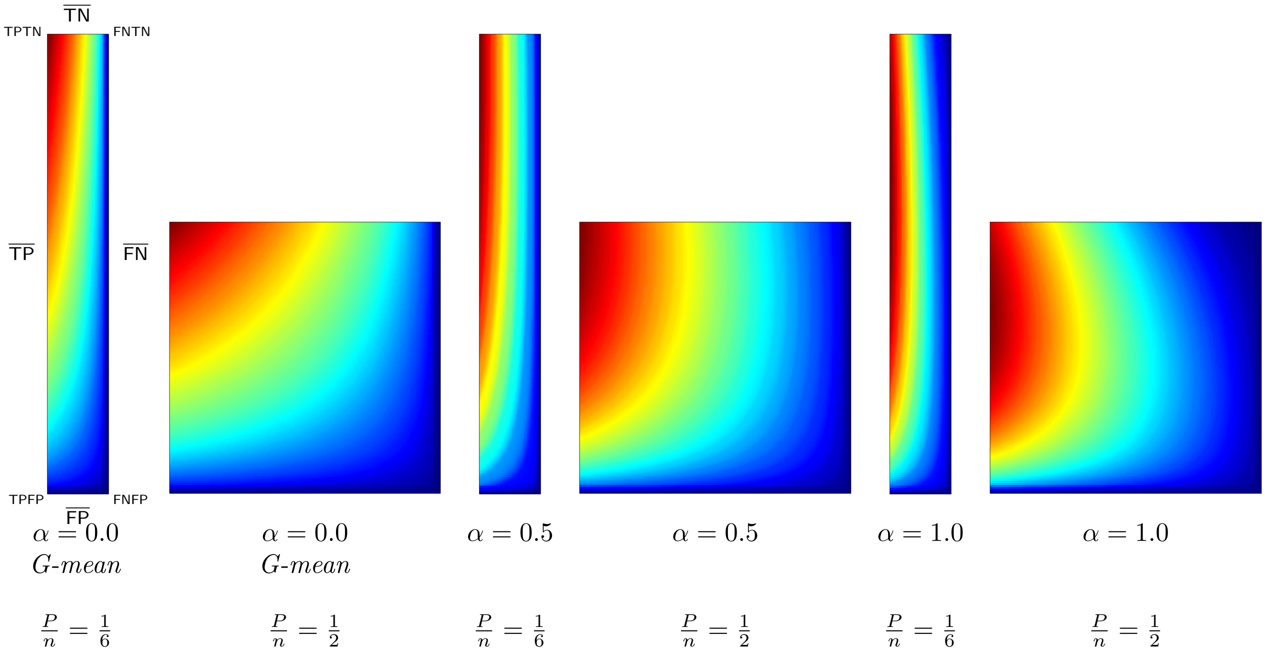


Fig. 16. Cross-sections of $IBA_\alpha(G\text{-}mean)$

Table 3 gathers the results concerning the ten devised properties for particular intervals of the α parameter implied by its border-line value.

Table 3: Properties of *G-mean* and its parametrizations; *: contains NaN (undefined value); †: NaN side, ^{s}: strong monotonicity

Measure	TPTN_{max}	$\overline{\mathsf{FN}}_{min}$	$\overline{\mathsf{FP}}_{min}$	$TP_{\nearrow}$	$TN_{\nearrow}$	$\overline{\mathsf{TN}}_{\neq max}$	$\overline{\mathsf{TP}}_{\neq max}$	*ACE*	*ACH*	*UnDefs*
$IBA_0(G\text{-}mean) = G\text{-}mean$	✓	✓	✓	✓	✓	✓	✓	✓	✓	TN–FP; TP–FN
$IBA_\alpha(G\text{-}mean)$, $\alpha \in (0, 1/3]$	✓	✓	✓	✓	✓	✓	✓	✓	×	TN–FP; TP–FN
$IBA_\alpha(G\text{-}mean)$, $\alpha \in (1/3, \infty)$	×	✓	✓	✓	×	✓	×	✓	×	TN–FP; TP–FN

In particular, the entries for *G-mean* and for $IBA_{\alpha\in(0,1/3]}(G\text{-}mean)$ state that such external parametrization eliminates the symmetry of handling both classes. It is the result of incorporating the class-asymmetric *Dom* component in the IBA_α parametrization procedure. The parametrized *G-mean* becomes slightly (as α does not exceed 1/3) more oriented towards the positive class (Fig. 16), and thus does not satisfy the *ACH* property any more. The behaviour of $IBA_\alpha(G\text{-}mean)$ changes more drastically when α exceeds 1/3 (Table 3, Fig. 16). On one hand, for $\alpha > 0$ one gets the much desired focus on the positive class, reflected by the *ACE* property (for any two corresponding points on sides $\overline{\mathsf{TP}}$ and $\overline{\mathsf{TN}}$, the value on side $\overline{\mathsf{TP}}$ is strictly greater than that on $\overline{\mathsf{TN}}$), however, for $\alpha > 1/3$ this comes with the inevitable cost of losing not only the above-mentioned *ACH* property, but also the $TN_{\nearrow}$ property (manifested by non-monotonic growth of the measure from $\overline{\mathsf{FP}}$ to $\overline{\mathsf{TN}}$). Additionally, for $\alpha > 1/3$ the maximal value of $IBA_\alpha(G\text{-}mean)$ drifts away from vertex TPTN. Thus, full recognition of both classes is no longer rewarded with the maximal measure value, violating the TPTN_{max} and $\overline{\mathsf{TP}}_{\neq max}$ properties. In this context, the usability of $IBA_\alpha(G\text{-}mean)$ for $\alpha > 1/3$ becomes questionable.

Issues raised in this section clearly demonstrate that the awareness of measure properties can strongly impact practical tasks of class imbalanced data. In this context, the proposed properties constitute valuable guidance for machine learning experts. Furthermore, the presented cases of visual inspection demonstrate how the barycentric visualization can aid researchers in these tasks.

7. Conclusions

In this paper, we examined classification performance measures applicable to class imbalanced problems. In order to support researchers in analysing and comparing various measures we proposed a new visualization technique that is based on the barycentric coordinate system, which properly reflects the multidimensional character of the measures' evaluation. We additionally provided an interactive tool that implements the technique in the form of a web application.

Independently, we put forward ten properties worth considering when choosing or designing measures for a given classification task. The purposefully defined properties verify aspects of measures that should be certainly realized before a measure is adopted for a given application, e.g. whether the measure possesses a class bias or not and, if so, in what degree this bias is preserved across the changing class proportions. This is especially true for measures of more complex formulae, which may include various forms of non-trivial parametrization.

To gain some practical insights, we analysed 22 popular classifier performance measures in terms of the ten properties. The selection included non-parametric, internally parametric as well as externally parametrized measures. The performed analyses led us to the identification of several important property changes in the measures. In particular, we have derived threshold values for selected properties of F_β and $IBA_\alpha(G\text{-}mean)$.

While various questions regarding measures can generally be answered after theoretical investigations (as demonstrated in the paper), many practical answers are often instantaneously provided by the proposed visualization technique. This form of analysis is especially useful for verifying the ranges of measure values and their changes across the domains of the measures. To aid researchers in these tasks, unlike simpler visualizations the proposed technique:

- provides general interpretations in terms of the four values of the two-class confusion matrix,
- involves exclusively linear, and thus easily interpretable, 4D → 3D transformations handling the four transformed values fully symmetrically,
- allows for analysing full ranges of measure values with respect to all possible combinations of confusion matrix entries,
- naturally illustrates the $TP + FN + FP + TN = n$ constraint, manifested in the shape of the space (i.e. tetrahedron),
- remains defined for all possible combinations of the matrix entries,
- admits multiple cross-sections with natural interpretations in terms of simple measures, e.g. horizontal cross-sections, which correspond to the proportion of actual classes (i.e. the positive (P/n) and the negative (N/n) class) and are thus especially well suited for analysis of imbalanced data.

Using the visualization technique, we demonstrated important differences in the ranges of measure values. Realizing such differences may prove fruitful in applications with more complex data characteristics, such as those of imbalanced classes, where selected measure properties are expected to be retained for wide ranges of changing data characteristics. It is worth stressing that it was not our intention to promote any single measure as the best, since the measure choice always finally depends on the user and the application at hand. Nevertheless, our visualization tool together with some of the theoretical results should support making such choices.

As future work concerning measure analyses, we plan to consider additional properties, such as measure related vector fields (e.g. gradients) as functions of the four arguments. Moreover, it would be interesting to analyse the effects of applying cost matrices to the visualized measures. Similarly, the effects of micro- and macro-averaging of binary measures in multi-class scenarios are worth studying. Finally, we plan to apply the visualization tool to particular classification results. Observe that in the process of classifier tuning the set of validation results often forms a trajectory, the analysis of which could prove useful. The course of such a trajectory may be illustrated using the introduced visualization technique, since a single validation result, being a confusion matrix, is a point in the tetrahedron. The visualization procedure could thus facilitate the classifier tuning process.

Acknowledgement

This research was partly supported by the FNP START scholarship (first author) and Institute of Computing Science Statutory Funds.

References

[1] R. Alaíz-Rodríguez, N. Japkowicz, P. E. Tischer, A Visualization-Based Exploratory Technique for Classifier Comparison with Respect to Multiple Metrics and Multiple Domains, in: Proc. 19th European Conf. Mach. Learn., Part II, 660–665, 2008.

[2] P. Baldi, S. Brunak, Y. Chauvin, C. Andersen, H. Nielsen, Assessing the Accuracy of Prediction Algorithms for Classification: An Overview, Bioinformatics 16 (2000) 412–424.

[3] M. Bekkar, H. Djemaa, A. Taklit, Evaluation Measures for Models Assessment Over Imbalanced Data Sets, Journal of Inform. Eng. and Appl. 3 (10) (2013) 27–38.

[4] J. Blaszczynski, J. Stefanowski, Neighbourhood sampling in bagging for imbalanced data, Neurocomputing 150 (2015) 529–542.

[5] P. Branco, L. Torgo, R. Ribeiro, A survey of predictive modeling under imbalanced distributions, ACM Comput Surv 49 (2) (2016) 31.

[6] D. Brzezinski, J. Stefanowski, Prequential AUC: Properties of the Area Under the ROC Curve for Data Streams with Concept Drift, Knowledge and Information Systems 52 (2) (2017) 531–562.

[7] R. Caruana, A. Niculescu-Mizil, Data Mining in Metric Space: An Empirical Analysis of Supervised Learning Performance Criteria, in: Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 69–78, 2004.

[8] E. Celotto, Visualizing the behavior and some symmetry properties of Bayesian confirmation measures, Data Min. Knowl. Discov. 31 (3) (2017) 739–773.

[9] J. Davis, M. Goadrich, The relationship between Precision-Recall and ROC curves, in: Proc. 23rd Int. Conf. Mach. Learn., 233–240, 2006.

[10] P. M. Domingos, A few useful things to know about machine learning, Commun. ACM 55 (10) (2012) 78–87.

[11] C. Drummond, R. C. Holte, Cost curves: An improved method for visualizing classifier performance, Machine Learning 65 (1) (2006) 95–130.

[12] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters 27 (8) (2006) 861–874.

[13] C. Ferri, J. Hernández-Orallo, R. Modroiu, An Experimental Comparison of Performance Measures for Classification, Pattern Recognit. Lett. 30 (1) (2009) 27–38.

[14] P. A. Flach, The Geometry of ROC Space: Understanding Machine Learning Metrics through ROC Isometrics, in: Proc. 20th Int. Conf. Mach. Learn., 194–201, 2003.

[15] M. S. Floater, Generalized barycentric coordinates and applications, Acta Numerica 24 (2015) 161–214.

[16] J. Fürnkranz, P. A. Flach, An Analysis of Rule Evaluation Metrics, in: Proc. 20th Int. Conf. Mach. Learn., 202–209, 2003.

[17] V. García, R. A. Mollineda, J. S. Sánchez, Index of Balanced Accuracy: A Performance Measure for Skewed Class Distributions, in: Proc. 4th Iberian Conf. Pattern Recognition Image Analysis, 441–448, 2009.

[18] V. García, R. A. Mollineda, J. S. Sánchez, Theoretical Analysis of a Performance Measure for Imbalanced Data, in: Proc. 20th Int. Conf. Pattern Recognition, 617–620, 2010.

[19] V. García, R. A. Mollineda, J. S. Sánchez, A bias correction function for classification performance assessment in two-class imbalanced problems, Knowledge-Based Systems 59 (2014) 66–74.

[20] J. Grzymala-Busse, J. Stefanowski, S. Wilk, A comparison of two approaches to data mining from imbalanced data, J Intell Manuf 16 (2005) 565–574.

[21] Q. Gu, L. Zhu, C. Z., Evaluation Measures of the Classification Performance of Imbalanced Data Set, in: Proc. ISICA, Springer, 461–471, 2009.

[22] H. He, E. A. Garcia, Learning from Imbalanced Data, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1263–1284.

[23] H. He, Y. Ma (Eds.), Imbalanced Learning: Foundations, Algorithms and Applications, IEEE - Wiley, 2013.

[24] J. Hernández-Orallo, P. A. Flach, C. F. Ramirez, Brier Curves: a New Cost-Based Visualisation of Classifier Performance, in: Proc. 28th Int. Conf. Mach. Learn., 585–592, 2011.

[25] B. Hu, W. Dong, A Study on Cost Behaviors of Binary Classification Measures in Class-imbalanced Problems, CoRR abs/1403.7100 .

[26] N. Japkowicz, Assessment Metrics for Imbalanced Learning, in: [23] (2013) 187–206.

[27] N. Japkowicz, M. Shah, Evaluating Learning Algorithms: A Classification Perspective, Cambridge University Press, ISBN 9780521196000, 2011.

[28] M. Kubat, R. Holte, S. Matwin, Machine Learning for the Detection of Oil Spills in Radar Images, Machine Learning Journal 30 (1998) 195–215.

[29] Y. Le Bras, P. Lenca, S. Lallich, Formal Framework for the Study of Algorithmic Properties of Objective Interestingness Measures, Data Mining: Foundations and Intelligent Paradigms: Volume 2: Statistical, Bayesian, Time Series and other Theoretical Aspects, Springer, 77–98, 2012.

[30] V. López, A. Fernández, S. García, V. Palade, F. Herrera, An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics, Inf. Sci. 250 (2013) 113–141.

[31] W. Mao, J. Wang, Z. Xue, An ELM-based model with sparse-weighting strategy for sequential data imbalance problem, Int. J. Machine Learning & Cybernetics 8 (4) (2017) 1333–1345.

[32] A. F. Moebius (Ed.), Der barycentrische calcul, Johann Ambrosius Barth, Leipzig, 1827.

[33] G. Piatetsky-Shapiro, Discovery, Analysis, and Presentation of Strong Rules, in: Knowledge Discovery in Databases, AAAI/MIT Press, 229–248, 1991.

[34] G. Piatetsky-Shapiro, B. M. Masand, Estimating Campaign Benefits and Modeling Lift, in: Proc. 5th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 185–193, 1999.

[35] D. M. Powers, What the F-measure Doesn't Measure: Features, Flaws, Fallacies And Fixes, CoRR abs/1503.06410 .

[36] F. J. Provost, T. Fawcett, R. Kohavi, The Case Against Accuracy Estimation for Comparing Induction Algorithms, in: Proc. 15th Int. Conf. Mach. Learn., 445–453, 1998.

[37] R. Ranawana, V. Palade, Optimized Precision - A New Measure for Classifier Performance Evaluation, in: Proc. IEEE Cong. on Evol. Computation, 16–21, 2006.

[38] Y. Sasaki, The truth of the F-measure, Tech. Rep., University of Manchester, URL http://www.cs.odu.edu/~mukka/cs795sum10dm/Lecturenotes/Day3/F-measure-YS-26Oct07.pdf, 2007.

[39] R. Susmaga, I. Szczęch, Can Interestingness Measures Be Usefully Visualized?, Int. J. Applied Math. Comp. Science 25 (2) (2015) 323–336.

[40] R. Susmaga, I. Szczęch, Visualization Support for the Analysis of Properties of Interestingness Measures, Bulletin of the Polish Academy of Sciences Technical Sciences 63 (1) (2015) 315–327.

[41] S. Vanderlooy, I. G. Sprinkhuizen-Kuyper, E. N. Smirnov, H. J. van den Herik, The ROC isometrics approach to construct reliable classifiers, Intell. Data Anal. 13 (1) (2009) 3–37.

[42] S. Wang, L. L. Minku, X. Yao, Resampling-Based Ensemble Methods for Online Class Imbalance Learning, IEEE Trans. Knowl. Data Eng. 27 (5) (2015) 1356–1368.

[43] J. Warren, S. Schaefer, A. N. Hirani, M. Desbrun, Barycentric coordinates for convex sets, Advances in Computational Mathematics 27 (3) (2007) 319–338.

[44] J. Zhai, S. Zhang, C. Wang, The classification of imbalanced large data sets based on MapReduce and ensemble of ELM classifiers, Int. J. Machine Learning & Cybernetics 8 (3) (2017) 1009–1017.